Investigating clinical outcomes in psychotic disorders using an electronic case register

Patel, Rashmi

Awarding institution:
King's College London

Investigating clinical outcomes in psychotic disorders using an electronic case register

Rashmi Patel

Thesis submitted for the degree of Doctor of Philosophy

April 2016

Department of Psychosis Studies

Institute of Psychiatry, Psychology & Neuroscience

King's College London

Adde parvum parvo magnus acervus erit

Abstract

Background:

Psychotic disorders have a lifetime prevalence of around 3% and cost the UK around £10 billion per year. One of the key challenges faced by clinicians in managing these disorders is that it is not possible to predict clinical outcomes or the course of illness. To date, research that has investigated clinical outcomes in psychosis has generally involved patient samples that are relatively modest in size and may be unrepresentative of the patients seen in everyday clinical practice. In many centres, routine clinical information is now recorded in the form of electronic health records (EHRs) rather than in paper case notes. While some of these data comprise responses to specific questions, at present most of the clinical data is still recorded in the form of unstructured free text entries. The large volume of free text means that it is not feasible to manually read through records to identify data of interest in a large sample of patients. However, automated information extraction methods such as natural language processing (NLP) offer the opportunity to quickly extract large volumes of meaningful data from free text EHRs and perform observational research studies in much larger samples than would be possible through direct participant recruitment.

Aims of this thesis:

To extract and analyse clinical data using NLP from a large electronic case register of patients with psychotic disorders to test the following hypotheses:

(i) In people with first episode psychosis (FEP), those who initially present to a service for people at high risk of psychosis will have better outcomes than those who present to conventional services.

(ii) Concurrent cannabis use in people with FEP is associated with poor clinical outcomes which are partly mediated by it reducing the effectiveness of treatment with antipsychotic medication

(iii) Negative symptoms are common in people with schizophrenia and are associated with worse clinical outcomes.

Methods (the same overall approach was used to test all hypotheses):

Dataset: South London and Maudsley NHS Trust (SLaM) Biomedical Research Centre (BRC) Case Register. The dataset comprises anonymised EHRs of over 250,000 people who have received mental healthcare from SLaM.

NLP development: The software package TextHunter was used. All sentences containing keywords relevant to the constructs investigated were extracted and used to develop NLP applications using a support vector machine learning (SVM) approach.

Outcomes: number of days spent in hospital and frequency of hospital admission.

Covariates: age, gender, ethnicity, marital status and diagnosis.

Statistical analysis: multivariable logistic, negative binomial, linear regression and mediation analysis using STATA.

Results:

FEP patients who initially presented to high-risk services (n=2,943): Presentation to a high-risk service was associated with 17 fewer days spent in hospital (95% CI -33.7, -0.3) and a lower frequency of admission (incidence rate ratio: 0.49, 0.39-0.61) in the 24 months following referral, as compared to patients who presented to conventional services.

Cannabis in FEP (n=2,026): Concurrent cannabis use was associated with increased frequency of hospital admission (incidence rate ratio 1.50, 1.25-1.80) and a greater number of days spent in hospital (B coefficient 35.1 days, 12.1-58.1). An increase in the number of unique antipsychotics prescribed to cannabis users mediated both an increased frequency of hospital admission (natural indirect effect: 1.09, 1.01-1.18; total effect: 1.50, 1.21-1.87) and a greater number of days spent in hospital (NIE: 17.9, 2.4-33.4; TE: 34.8, 11.6-58.1).

Negative symptoms in chronic schizophrenia (n=7,678): 55.7% of people with schizophrenia had at least one negative symptom documented. Among patients with schizophrenia, negative symptoms were associated with increased likelihood of hospital admission (odds ratio 1.24, 95% CI 1.10-1.39), re-admission (1.58, 1.28-1.95) and length of stay (B coefficient 20.5, 7.6-33.5).

Conclusions:

EHR data can be used to investigate associations between variables assessed during routine care and clinical outcomes in patient samples that are much larger than can be recruited to conventional research studies. Moreover, the specific findings obtained using this approach have a number of implications for healthcare service delivery. First, the finding that engaging first episode patients in the prodromal phase is associated with better outcomes indicates that contact at this stage may not only reduce the risk of developing psychosis, but also improve outcomes in those at high risk who subsequently become psychotic. Secondly, the finding that both cannabis use and negative symptoms in patients with psychosis are independently linked to significantly poorer clinical outcomes highlights the need for the development of effective treatments to reduce cannabis use and ameliorate negative symptoms.

Table of contents

Acknowledgments

I would like to thank Professor Philip McGuire for his unwavering guidance and supervision throughout my PhD and clinical academic career. I first met Philip during a Special Study Module at the Institute of Psychiatry in my final year at medical school. My experience kindled an ongoing interest in Academic Psychiatry leading me to pursue the work presented in this thesis. I am extremely grateful for his longstanding support and for providing me with the opportunity to pursue a clinical research career in Psychosis Studies. I am also very grateful to Professor Robert Stewart for his ongoing support as my second supervisor, for patiently reviewing my research manuscripts, and for encouraging me to freely collaborate with the informaticians and researchers in the SLaM BRC Nucleus team who have supported my research.

I would like to thank all my colleagues in the Department of Psychosis Studies and the SLaM BRC Nucleus supporting my research. In particular, I would like to thank Paolo Fusar-Poli, Sagnik Bhattacharyya, Richard Jackson, Michael Ball, Hitesh Shetty, and Matthew Broadbent for their support in undertaking the research studies presented in this thesis.

I am very grateful to Professor Chris O'Callaghan for his wisdom and guidance as my academic mentor, helping me to navigate through the uncharted and, at times, turbulent path of clinical academic training. I am also grateful to Professor Chris Pugh, Dr Denise Best and other colleagues in the Oxford University Clinical Academic Graduate School (OUCAGS) who have supported my predoctoral career as an academic clinical fellow.

I would like to thank the Medical Research Council for awarding me a Clinical Research Training Fellowship to undertake the work leading to this thesis.

Finally, I would like to thank all the patients, their carers and mental healthcare professionals in the South London and Maudsley NHS Trust who were ultimately responsible for obtaining the enormous volumes of clinical data I have analysed in my research. Without them, none of these studies would have been possible.

Structure of thesis

This is a thesis incorporating four published journal articles. These articles are incorporated in chapters 3, 4, 5 and 6. A summary of the structure of each chapter is provided below.

Chapter 1: an introductory chapter which first reviews the literature on different stages of psychotic disorders and previous studies investigating clinical outcomes in these disorders. This is followed by some background on the use of electronic case registers and natural language processing for clinical research and concludes with a statement of the aims and objectives of the work reported in this thesis.

Chapter 2: a detailed description of the methods employed in the research. This includes a description of the electronic health record system, the electronic case register, the database methods used to interrogate the case register and the natural language processing methods employed to obtain clinical data from unstructured text.

Chapters 3, 4 and 5: these chapters incorporate the journal articles (including supplementary material) which report the original research undertaken in this thesis. This includes the background to each study, methods employed, results and discussion of findings. Each chapter is accompanied by a supplementary methods section to provide further detail on methods which are unique to each study where there was insufficient space available to describe these in the main text.

Chapter 6: this comprises a co-authored article on the TextHunter software used for natural language processing to facilitate the description of the natural language processing methods in chapter 2 and section 4.23.

Chapter 7: this chapter summarises and brings together key findings from each of the studies reported in this thesis followed by a discussion on the strengths and limitations of the studies. This is followed by a discussion of the implications of the findings of these studies on future research.

References: references which are quoted outside of sections 3.1, 4.1, 5.1 and 6.1 are provided here in the Vancouver format.

Tables and figures: with the exception of the journal articles incorporated in chapters 3, 4, 5, and 6, tables and figures for the main thesis are provided in-line with the text and referenced using a numbering system in which the first number refers to the chapter in which the table or figure occurs. The numbering of tables and figures in chapters 3, 4, 5 and 6 refers to tables and figures within each of these chapters and not to the tables and figures presented in the remainder of the thesis.

Contribution statement

Chapters 1 and 2: the content of these chapters are entirely the work of Rashmi Patel, with the exception of Figure 2a, a screenshot created by the Department of Computer Science, University of Sheffield, which is used for illustration purposes to facilitate description of the method in section 2.61. Specific support for the methods described in chapter 2 was provided by Hitesh Shetty and Matthew Broadbent (section 2.4), and Richard Jackson (section 2.6).

Chapter 3: Paolo Fusar-Poli, Covadonga Díaz-Caneja and Rashmi Patel designed the study. Data collection for patients presenting to the high risk clinical service was performed by Covadonga Díaz-Caneja. Data collection for patients presenting to conventional mental health services was performed by Rashmi Patel with support from Hitesh Shetty. Statistical analysis was performed by Rashmi Patel. Drafting of the manuscript was performed by Paolo Fusar-Poli, Covadonga Díaz-Caneja and Rashmi Patel. Overall study supervision was provided by Paolo Fusar-Poli and Philip McGuire. Critical review of the manuscript was performed by all listed authors in section 3.1. Paolo Fusar-Poli, Covadonga Díaz-Caneja and Rashmi Patel were joint first authors. The description of supplementary methods (section 3.2) is entirely the work of Rashmi Patel.

Chapter 4: The study was conceived by Sagnik Bhattacharyya. The data extraction was led by Rashmi Patel with support from Richard Jackson, Michael Ball and Hitesh Shetty, supervised by Matthew Broadbent and Robert Stewart. Statistical analyses were carried out by Rashmi Patel and reporting of findings by Rashmi Patel and Robin Wilson, supervised by Philip McGuire and Sagnik Bhattacharyya. All authors contributed to manuscript preparation and approved the final version. The description of supplementary methods (section 4.2) is entirely the work of Rashmi Patel.

Chapter 5: The study was conceived by Robert Stewart and Nadia Foskett. The CRIS-NSS product development was led by Richard Jackson with significant input from Matthew Broadbent, Genevieve Gorrell, Caroline Johnston, Angus Roberts and Hitesh Shetty. Initial analyses were carried out by Robert Stewart, Chin-Kuo Chang and Richard D Hayes. Final analyses and reporting of findings were led by Rashmi Patel and Nishamali Jayatilleke, supervised by Robert Stewart and Philip McGuire. All authors contributed to manuscript preparation and approved the final version (section 5.1). The description of supplementary methods (section 5.2) is entirely the work of Rashmi Patel with the exception of the SQL scripts quoted in section 5.21 and 5.22, which are reproduced with the kind permission of Hitesh Shetty, who performed the SQL data extraction for this study.

Chapter 6: The TextHunter manuscript was co-authored with Richard Jackson et al. Specifically, Rashmi Patel contributed to the development of the cannabis NLP application, which was a test case for the TextHunter application, and reviewed and edited the final draft of the manuscript.

Chapter 7: the content of this chapter is entirely the work of Rashmi Patel.

Overall supervision: I am grateful to Philip McGuire and Robert Stewart for their overall supervision to the work presented in this thesis.

Novel contribution of this thesis to existing literature

Knowledge

Chapter 3: around one third of people who present to clinical services designed to assess and treat people at high-risk of developing psychosis are already experiencing their first episode of psychosis (FEP). Prior to this study, the clinical outcomes of these individuals were unknown. This study adds to the literature by demonstrating that people with FEP who initially present to high-risk clinical services have better clinical outcomes than those who present to conventional mental health services. These findings have important implications for mental healthcare service delivery as they suggest that the provision of high-risk clinical services may help to improve detection and treatment of people with an already established psychotic disorder (as well as people who are at high risk, but not yet psychotic).

Chapter 4: existing research suggests that exposure to cannabis is associated with an increased risk of developing a psychotic disorder. However, there is less published literature investigating the association of cannabis with clinical outcomes in established FEP and many of these studies are based on relatively modest sample sizes drawn from research cohorts. Furthermore, the underlying mechanisms by which cannabis is associated with poor clinical outcomes are unclear. This study adds to the literature by demonstrating a clear association between cannabis use and substantially poorer clinical outcomes following first presentation to mental health services with a psychotic disorder. Moreover, the study involved a large sample of patients that is representative of the overall patient population receiving mental healthcare for FEP. In addition, this is the first study to investigate a possible explanation of poorer clinical outcomes related to a failure of antipsychotic treatment as demonstrated through a mediation analysis.

Chapter 5: previous studies suggest that negative symptoms account for a substantial proportion of morbidity and disability in people with schizophrenia. However, these findings are largely based on the application of specialised symptom inventories applied in research studies which may not be representative of everyday clinical practice. This study adds to the literature by demonstrating that negative symptoms are highly prevalent in people receiving routine mental healthcare for schizophrenia, and that they are associated with particularly poor clinical outcomes. These findings highlight the importance of assessing patients with schizophrenia for negative symptomatology, and the need for better treatments for these symptoms.

Methods

Section 2.4: the methods described in this section explain the use of Microsoft Structured Query Language to transform and recode clinical data derived from anonymised electronic mental health records. The methods described in this section may be applied to other studies on the same dataset or similar datasets using Microsoft SQL.

Section 2.51: this section outlines a novel method to recode mental disorder diagnoses extracted from the unstructured free text of electronic health records. This method may be applied to similar datasets using the SQL scripts outlined in this section.

Section 2.52: this section outlines a novel method to recode antipsychotic medication data extracted from the unstructured free text of electronic health records. This method may be applied to similar datasets using the SQL scripts outlined in this section.

Section 2.6: the methods described in this section describe novel text mining methods which may be used to extract clinical information from unstructured free text data from electronic health records. The principles described may be applied to any dataset of free text clinical data in any area of Medicine.

1. Introduction

1.1 Epidemiology and natural history of psychotic disorders

Psychotic disorders have a lifetime prevalence of around 3%[1] and cost the UK around £10 billion per year.[2] They can include a wide range of clinical disorders including schizophrenia, schizoaffective disorder, bipolar disorder and psychotic depression, depending on how psychotic disorders are defined. One of the challenges faced by clinicians in treating individuals with psychotic disorders is that it is currently not possible, at an individual level, to predict the prognosis or which treatments are likely to be of benefit. This may reflect the fact that the conventional DSM/ICD diagnostic categories of different psychotic disorders have overlapping genetic, neurobiological and clinical features.[3] Furthermore, there is considerable heterogeneity *within* each diagnostic category, such that the diagnostic label provides little information about the course or outcome in an individual patient.[4]

The onset of psychotic disorders is preceded by a prodromal period, during which an affected individual may experience 'attenuated' and/or brief psychotic symptoms in the context of a marked functional decline.[5] Over the last two decades, specialised mental health services (hereafter referred to as *high risk services*) have been developed to identify and offer treatment to people who have a very high risk of developing FEP.[5,6] High risk services identify people who present with clinical features (known as the at-risk-mental state, or ARMS) that have been shown to be associated with a risk of developing psychosis of 20-35% in the next 3 years. The clinical criteria include any one of the following, in association with functional decline and distress:[5]

(i) Attenuated Psychotic Symptoms (APS): where an individual experiences psychotic symptoms but at a lower intensity or frequency than that of a psychotic episode.

(ii) Brief Limited Intermittent Psychosis (BLIP): a self-remitting episode of psychosis which resolves within 7 days.

(iii) Genetic Risk and Deterioration Syndrome (GRD): genetic risk as indexed by a positive family history of psychosis.

However, only around 35% of people presenting to high risk services meet the criteria for ARMS. 32% of people referred to these services are found to already have developed FEP at presentation.[6] Little is known about the difference in clinical outcomes between FEP patients whose first contact with services is through high risk teams and those who present to conventional mental health services.

The onset and progression of a psychotic illness following first presentation is varied. Estimates of the proportion of patients who achieve remission of symptoms within 2 years of first presentation vary between around 15%[7] and 50%.[8] It is thought that the majority of patients (77%) experience a sustained period of remission of at least 6 months, but a significant minority go on to develop more serious illness characterised by frequent hospital admissions.[8] Furthermore, around 20-30% of patients with schizophrenia do not respond to first-line antipsychotic treatment and may experience ongoing residual psychotic symptoms.[9] It is thought that a number of clinical and sociodemographic factors may predict better outcome following FEP. These include older age at onset, female gender, and presentation with an affective psychosis (e.g. schizoaffective disorder, mania or psychotic depression).[10] However, at present, there is no way to reliably predict response to treatment or prognosis in this group at individual patient level.

1.2 Negative symptoms in psychotic disorders

Factor analytic studies suggest that psychotic disorders are characterised by five different symptom dimensions. These include positive symptoms, negative symptoms, mania, depression and cognitive symptoms[11–14]. Regression analyses have demonstrated that the presence and severity of these symptom domains at presentation predict functional outcomes 10 years later[15]. Positive psychotic symptoms have been the major focus of therapeutic targets for pharmacological and psychological interventions. However, there is increasing evidence that negative symptoms, which have been relatively neglected as targets for therapy, predict poorer clinical outcomes in patients with psychosis[16]. These symptoms include flattened affect, avolition, apathy, emotional withdrawal, social withdrawal and diminished expressivity. It is thought that they may account for a greater burden of illness than any other feature of psychotic disorders [17].

The presence of negative symptoms in psychotic disorders predicts poorer long term outcomes, with poorer psychosocial functioning and reduced likelihood of remission.[16] These associations are found in individuals experiencing their first episode of illness[18–20] as well as patients with a chronic disorder,[21] and in adolescent[22,23] as well as adult and older adult patients.[24] Factor analytic studies indicate that negative symptoms comprise two discrete subdomains, one relating to avolition/apathy and the other to diminished expressivity.[25] Recent evidence suggests that the former subdomain is particularly associated with poor functional outcomes in psychosis.[26]

The aetiology and pathophysiology of negative symptoms are still unclear.[27] However, neuroimaging studies in schizophrenia suggest that negative symptoms are associated with alterations in the structure[28,29] and function[30] of the prefrontal cortex and the striatum, and with elevated cingulate glutamate function.[31]

In contrast to positive symptoms, there are currently no established treatments for negative symptoms, partly because the underlying pathophysiology is poorly characterised.[32] Interventional studies have investigated the role of antidepressants,[33,34] second generation antipsychotics,[35,36] psychological therapies,[37] folate and vitamin B12,[38] modafinil,[39] non-steroidal anti-inflammatories[40] and transcranial magnetic stimulation[41] in treating individuals with negative symptoms in psychosis with varying degrees of success. More recently, there has been growing interest in the use of novel glutamatergic modifiers to improve negative symptoms.[32,42]

Instruments such as the Brief Negative Symptoms Scale (BNSS),[43] and the Scale for the Assessment of Negative Symptoms (SANS)[44] have been shown to reliably elicit negative symptomatology. However, the application of these instruments in standard clinical practice is limited by their complexity, the time required to complete an assessment, and the need for the raters to be trained in their use. For this reason, there is a need to develop methods for ascertaining negative symptoms that are reliable but practical to implement in clinical practice in order to better characterise the association of negative symptoms on clinical outcomes.

1.3 Cannabis use in psychotic disorders

Cannabis is the most widely used illicit substance, with almost 200 million users worldwide.[45] In England, the overall prevalence of cannabis use in 2007 was 7.5%, with a history of lifetime use reported among 27.8% of men and 18.6% of women.[46] Its use in adolescents and young adults is associated with reduced educational achievement and social and occupational functioning.[47–49] There is a growing body of evidence that cannabis use is associated with the development of psychotic disorders,[50,51] particularly when used regularly from early adolescence.[52] High potency cannabis ('skunk') has been reported to be associated with a population attributable fraction of 24% for developing FEP.[53] It is thought that the increased risk of FEP among individuals who use cannabis may be associated to genetic predisposition or vulnerability during neural development.[54,55]

Among people with psychotic disorders, cannabis use is associated with a younger age of onset[56] and a longer duration of untreated psychosis (DUP).[57] In people with chronic psychosis, cannabis use is associated with more severe psychotic symptoms,[51] worse medication adherence,[58] more frequent relapse,[59] suicidal behaviour[60] and worse overall functioning.[58] However, much less is known about the association of cannabis use with clinical outcomes in patients in the early stages of psychosis, particularly with respect to its potential impact on the effectiveness of treatment, as previous studies have been limited by relatively small sample size and duration of follow-up.[56,58,61–64]

1.4 Electronic case registers

An electronic case register is a dataset which contains information on individuals receiving healthcare within a defined population.[65] Such datasets may originally be developed for specific clinical reasons, such as to facilitate management of healthcare services or for public health surveillance. However, the same data may be also used for secondary purposes such as audit and clinical research.[66] In the UK, several electronic case registers have been developed for this purpose in a wide range of medical disciplines including Renal Medicine,[67] Oncology[68] and Primary Care.[69] Outside the UK, some national databases such as the Danish Psychiatric Central Research Register (PCRR) have been specifically designed to facilitate mental health research.[70] The analysis of large scale case registers has contributed to a better understanding of the epidemiology of schizophrenia[71] and suicide[72], as well as adverse physical health outcomes in people with psychotic disorders.[73] Such datasets benefit from sample sizes that are orders of magnitude larger than those typically used in conventional research studies, thus improving their statistical power. However, one of the disadvantages is a relative lack of rich clinical data which would be obtained through face-to-face assessment of participants recruited to a research study.[74] Historically, this has limited the scope of mental health research involving case registers to studies of incidence and prevalence of disorders and the overall association of sociodemographic factors and healthcare services with clinical outcomes.

Figure 1: Psychosis case register research OVID MEDLINE search

Search History (8 searches) (close)

View Saved

#	Searches	Results	Search Type	Actions
1	(psychotic or psychosis or schizo*).ti.	89956	Advanced	Display More »
2	(Directory/ or dataset.mp. or database.mp. or register.mp. or registry.mp.) not cochrane.mp.	268844	Advanced	Display More »
3	Treatment Outcome/ or outcomes.mp.	1083201	Advanced	Display More »
4	1 and 2 and 3	178	Advanced	Display More »
5	from 4 keep 12, 22-23, 25, 27, 31, 34... Select out publications listed as reviews or interventional trials	52	Advanced	Display More »
6	4 not 5	126	Advanced	Display More »
7	remove duplicates from 6	124	Advanced	Display More »
8	from 7 keep 3, 5-9, 15, 17-19, 21-23, 27... Manually select relevant publications	93	Advanced	Display More »

Remove Selected | Save Selected | Combine selections with: And Or | RSS

Save Search History

1. Introduction

A literature search of the MEDLINE database using OVID Gateway[75] (search strategy outlined in Figure 1) identified 93 published articles (up to 16th August 2015) containing key words related to clinical outcomes in psychotic disorders and case registers. The sources of data for these studies varied between national statutory case registers designed to record data across the entire population of a country (Hospital Episode Statistics, UK[76]; Central Psychiatric Research Register, Denmark[77]; National Health Insurance Research Database, Taiwan[78–80]) to purpose-built registers on specific clinical populations (Tianjin Urban Employee Basic Medical Insurance (UEBMI) database, China[81]; Medicaid MarketScan database, USA[82,83]) and research registers to investigate specific exposures and clinical outcomes (International Observational Registry on Schizophrenia (InORS)[84]; Electronic Schizophrenia Treatment Adherence Registry (eSTAR)[85,86]; Early Intervention Service clinical database, Hong Kong[87]).

The outcomes that can be analysed using case registers depend on the type of data available within the case register being investigated. Largescale national case registers typically permit analysis of the exposures and clinical outcomes that are routinely recorded in all members of the population. These may include sociodemographic factors, mortality, hospitalisation and cost of healthcare.[81,82,88] More specialised case registers which include clinical data from healthcare services or from research studies permit the analysis of more specific exposures and outcomes. These include the effects of medication on outcomes[84,85,89,90] as well as the impact of psychotic disorders on physical health outcomes.[79,91] Some studies benefit from the combination of different case registers using a data linkage. This involves using a common participant identifier (such as name and date of birth or social security/healthcare ID number) to identify data from two different registers at individual participant level. One example includes a linkage between the Danish IVF register and the Danish Central Psychiatric Research Register to investigate differences in assisted reproductive technology outcome in people with schizophrenia.[92]

In some electronic case registers (such as the General Practice Research Database[69]), data are obtained directly from electronic health records (EHRs). The primary purpose of an EHR is to allow clinicians to document clinical information on individual patients in the course of providing healthcare within defined clinical services. These data may include clinical assessments, notes on progress during an episode of treatment, and correspondence between different healthcare professionals.[93] More recently, electronic case registers have been developed using EHR data from mental healthcare services. One example of this is the Biomedical Research Council (BRC) South London and Maudsley NHS Trust (SLaM) Case Register which comprises EHR data from over 250,000 patients receiving care from a large provider of mental health services in South London.[93] Pooled data from the SLaM BRC Case Register has been used to undertake a wide range of epidemiological studies[94–97].

EHRs typically include structured text fields (including patient demographics and structured forms for recording clinical data) as well as unstructured free text from progress notes recorded during patient/clinician interaction. Analysis of mental health EHRs suggests that the majority of rich clinical data (such as data on clinical presentation, diagnosis and treatments) are documented in unstructured free text, rather than in structured items.[98] However, extracting data from unstructured text can be time-consuming if this entails reading through long pieces of text in a large sample of patients. Fortunately, automated information extraction methods that facilitate obtaining data from large volumes of unstructured text have recently been developed. These are described further below.

1.5 Natural language processing

Natural language processing (NLP) is an information extraction method which allows useful information to be identified from unstructured free text documents.[99] NLP involves the use of an algorithm based on a series of rules to identify constructs of interest. Machine learning techniques can also be employed on volumes of text which have previously been classified by a human annotator to derive an automated algorithm to identify a construct of interest.[100]

The recent move to storing clinical records in electronic format has provided the opportunity to apply NLP to unstructured free text within EHRs. NLP has been previously applied to EHR data to investigate post-operative complications,[101] adverse outcomes from prescribed drugs,[102] clinical outcomes in depression.[103] The use of NLP thus permits rapid extraction of clinical constructs from large volumes of text. The technique also permits the extraction of subtle clinical constructs (such as symptoms of mental disorders) which are not routinely documented in structured fields within the electronic health record and so would otherwise be unavailable for analysis. NLP methods have been applied to EHR data from the SLaM BRC Case Register to extract a number of clinical constructs including smoking status,[104] Mini Mental State Examination (MMSE) scores[105] and symptoms of mood instability.[106]

1.6 Aims and objectives

In this thesis, I have applied data extraction techniques (including NLP) to data from the SLaM BRC Case Register to investigate the three hypotheses related to clinical outcomes in psychotic disorders described in sections 1.1, 1.2 and 1.3:

(i) Chapter 3: People with FEP who present to a clinical service for people at high risk of psychosis will have better outcomes than those who present to conventional mental health services.

(ii) Chapter 4: Cannabis use in people with FEP will be associated with poor clinical outcomes, and these are partly mediated by an adverse impact on the effect of antipsychotic therapy.

(iii) Chapter 5: Negative symptoms are common in people with chronic schizophrenia and are associated with particularly poor clinical outcomes.

2. Methods

2.1 Source of clinical data

All clinical data analysed and presented in this thesis were obtained from the South London and Maudsley NHS Foundation Trust (SLaM). The UK National Health Service (NHS) is organised into primary, secondary and tertiary healthcare services.[107] These services are managed by clinical commissioning groups (CCGs) who determine the structure and funding of healthcare services in a given geographical area. Primary care includes a range of community healthcare services (including general practices, walk-in centres and pharmacists) which UK residents may access to obtain assessment and treatment for common healthcare problems. If a patient's illness is serious or complex and cannot be effectively managed in primary care, they may be referred to a secondary healthcare service. These services are more specialised and include acute general hospital trusts and mental healthcare trusts that typically serve a catchment area of one or more geographical boroughs. Even more specialised tertiary healthcare services (for complex disorders requiring assessment and treatment in a world-leading specialist centre) may cover a wider catchment area of several counties or the whole country.[107] SLaM is a large provider of secondary mental healthcare covering a catchment population of around 1.2 million people residing in the boroughs of Lambeth, Southwark, Lewisham and Croydon in South London. SLaM also provides specialised inpatient and community mental health services to people with complex mental disorders through tertiary mental healthcare services.[108]

Entry into a SLaM service typically requires a referral to be received by one of its community team or hospital wards. Referrals may be initiated by general practitioners (GPs) who are medically qualified practitioners providing primary healthcare services in the community. Referrals may also be initiated by medical specialists in acute general hospitals, other mental health services, private healthcare services (i.e. non-NHS) and the criminal justice system. Some SLaM services also accept direct referrals from patients and their carers.[6]

2. Methods

Clinical services in SLaM are largely structured around the following service streams: Child and Adolescent Mental Health Services (CAMHS – up to 18 years); General Adult services (18-65 years); Older Adult services (older than 65 years). These service streams include Community Mental Health Teams (CMHTs) and inpatient clinical services. CMHTs provide community mental healthcare to patients with a mental disorder within a defined geographical catchment area, typically linked to the geographical catchment area of GPs in the same area. CMHT care is delivered by a multidisciplinary team including psychiatrists, psychologists, care co-ordinators (who may be qualified in mental health nursing or social care), pharmacists and other allied healthcare professionals. Inpatient clinical services provide high intensity care to patients who present with serious mental illness which necessitates hospital admission for acute care or rehabilitation. In addition to CMHTs and inpatient services, SLaM provides more specialised services which include forensic mental healthcare (for patients receiving compulsory mental healthcare following a criminal conviction) and home treatment teams (providing high intensity care in the community with the aim of avoiding hospital admission).

Since 2009, healthcare services in South London have been restructured into Clinical Academic Groups (CAGs) which are part of the King's Health Partners Academic Health Sciences Centre (AHSC).[109] The CAGs were developed as part of the AHSC initiative in order to better integrate the provision of clinical care in specialist NHS services with clinical research. The CAGs providing General Adult mental healthcare in South London include Addictions; Behavioural and Developmental; Mood, Anxiety and Personality; Psychological Medicine; Psychosis. Within the Psychosis CAG, clinical services are structured around four service lines based on different stages of illness: Early Intervention,[110] for patients experiencing their first episode of psychosis; Community, providing care to people with stable, chronic psychotic disorders in the community; Promoting Recovery, for patients recovering from an acute or chronic psychotic episode, typically based in the community;[111] Complex Care, a service which offers rehabilitation and recovery interventions in the community or on long-stay inpatient ward for patients with a long standing, severe psychotic disorder and possible comorbid psychiatric disorders; Acute Care, a hospital-based inpatient service for patients experiencing an acute episode or relapse of psychosis.

2.2 Electronic Patient Journey System

Clinical records in SLaM are documented in the electronic Patient Journey System (ePJS), an electronic health record (EHR) which is used by mental healthcare professionals to document routine clinical information in the course of delivering patient care.[93] Prior to the implementation of ePJS, clinical records were largely documented on paper, either in the form of written notes or printed correspondence and hospital discharge summaries. The implementation of ePJS in October 2005 led to all subsequent documentation being recorded electronically. By April 2006, all SLaM services (apart from the Addictions service, where ePJS was implemented in 2008) had migrated to exclusively using ePJS.

Clinical data may be documented into ePJS in structured fields or unstructured fields. Structured fields are defined as those in which entry of data is constrained in some way. These constraints may include entering a particular type of data (integer, date, time, string) or selecting from a list of predefined options (e.g. male vs female). Structured fields are designed to standardise the documentation of certain types of clinical data in order to facilitate their documentation, presentation and aggregation. In contrast, unstructured fields have no constraints on the type of data that are stored and permit users to type and store their own composition of free text. However, some unstructured fields may have constraints on the length of entry.

The fields in ePJS are organised in forms. Each form is designed store a collection of relevant fields to document a particular set of clinical data (e.g. demographic information, diagnosis, standardised clinical assessment) or a particular clinical event (e.g. an outpatient consultation, inpatient ward round, home visit). Every patient receiving care in SLaM has an ePJS account that is created when they are first referred to a SLaM clinical service. Within each patient's ePJS account, clinicians may create and edit forms throughout the course of a patient's clinical care. This collection of forms may be edited and viewed by clinicians in a graphical user interface (GUI) through a web browser on the SLaM intranet, an internal computer network accessible only to authenticated users and computer devices in SLaM. In this way, the clinical record of an individual patient may be built through multiple clinical encounters over time. Clinicians can thus view historical data in patients' records in order to determine ongoing clinical management. Multiple patient records may also be reviewed together for the purposes of managing and auditing the performance of a clinical team.

2.3 SLaM Biomedical Research Centre (BRC) Case Register

The SLaM BRC Case Register is an electronic mental health case register derived from clinical records stored in ePJS.[93] This case register is populated via an automated pipeline which extracts and anonymised clinical data from ePJS and stores it in a Microsoft Structured Query Language (SQL) Server database.[112] These data may then be accessed for the purposes of clinical research and audit using the methods described subsequently in section 2.4. In essence, the SLaM BRC Case Register is a searchable database of the clinical records of all patients who have received healthcare in SLaM that has been documented in ePJS (excluding patients who opted out of being included in the case register) and enables researchers to perform large-scale epidemiological analyses of specific cohorts to investigate the incidence and prevalence of mental disorders and associations of exposures of interest with clinical outcomes. With the exception of redacted patient identifiers, the data in the BRC Case Register are identical to those recorded in ePJS by mental healthcare professionals.

The SLaM BRC Case Register and CRIS have received ethical approval from the Oxfordshire Research Ethics Committee C (08/H0606/71+5) as an anonymised dataset for mental health research.[113] A patient-led oversight committee provides governance for all projects conducted using these data. A robust firewall and data security framework governs access to clinical data from the case register and only approved researchers are permitted to access data from the case register.[113] Patients who do not wish their anonymised data to be included in the case register may opt-out at any time. To date, fewer than 10 patients out of over 250,000 have opted out.

The automated method for extracting data from ePJS into the BRC Case Register involves a pattern matching approach to determine pieces of data which indicate patient identifiers from fields known to contain identifying information in ePJS (e.g. name, address, date of birth).[113] For each patient, a search is run on their ePJS record to find instances where identifiers have been documented. These are then replaced with a string of characters to indicate that they have been redacted (e.g. "ZZZZZZ"). The de-identified data are then copied into a Microsoft SQL Server database. Data may be obtained from the SQL database using the Clinical Record Interactive Search tool described in section 2.4.

The BRC Case Register data are represented in the SQL database in a series of tables which correspond to the various forms which are available for completion in ePJS by mental healthcare professionals. As patient identifiers are removed from records in the case register, an identification number (known as the BRCID) is generated for each patient. Every table in the SQL database includes a column for the BRCID in order to allow for data for the same patient from different forms to be analysed together.

2.4 Clinical Record Interactive Search tool

Data are accessed from the BRC Case Register using the Clinical Record Interactive Search tool (CRIS).[93] The CRIS software includes two interfaces: a web-based search engine powered by the Microsoft FAST system which allows users to perform key word searches in different parts of the case register, and a SQL Server Management Studio[112] interface which allows users to perform complex data queries to allow for data transposition, recoding and joining between different sources. All data obtained in the studies presented in this thesis were obtained using the SQL Server Management Studio interface.

The BRC Case Register is a relational database in which data stored in tables within the database may be joined together by means of a common identifier represented in multiple tables. In the BRC Case Register, two key identifiers are used to join data between different sources:

(i) BRCID: this is the unique identifier for each patient and allows data from different parts of the case register to be joined together at the patient level.

(ii) CN_Doc_ID: this is a unique identifier for each form within a given table.

In this way, each BRCID (i.e. each patient) may have many CN_Doc_IDs (i.e. forms or documents stored within their clinical record). By writing specific queries, it is possible to join data from different tables on a single BRCID in order to produce an output table for a cohort of patients in which each patient is represented as a single row in the table. The data may then be analysed using statistical software (described in further detail in section 2.7).

The method for extracting data involves writing a series of commands in a SQL script that select, transform and recode data from source tables within the database. SQL queries are not case sensitive (i.e. commands may be upper or lower case) or white-space sensitive (i.e. additional line breaks and spaces can be placed between separate commands). SQL queries may include "statements" which instruct the database to perform a particular query or command and "functions" which arithmetically transform the data in some way. Complex SQL functions may be stored in a "store procedure" which may be requested within a SQL script. Each command (statement, function or store procedure) is separated by a comma. An initial query typically includes the following commands:

(i) USE: this specifies the name of the database to search

(ii) SELECT: this specifies which variables to obtain. By default, the variable is given the same name as the variable from the source table. However, another variable name can be specified to distinguish it from the variable in the source table.

(iii) FROM: this specified the source table from which the variables are obtained

By default, the names of databases, tables and columns of tables are placed in parentheses of square brackets (i.e. "[]"). This is to allow for special characters such as spaces or punctuation marks to be used in database, table and column names. However, if the database/table/column name only contains alphanumeric characters, the square brackets are not necessary. For example, the following describes a table in the SQLCRIS_User database:

[SQLCRIS_User].[RPatel].[rp_Test_cohort]

However, as neither the database nor table name the following nomenclature would also be acceptable:

SQLCRIS_User.RPatel.rp_Test_cohort

The data obtained from a query may be saved into a new table (from which a subsequent query may be run) by using the INTO command. Alternatively, the SQL query may be saved as a "view" within the database. When queries are run on a view, the SQL query which generated the view is dynamically run to obtain the data as if it were already stored in a table.

In sections 2.41 to 2.45 I outline the main SQL methods which are common to all the studies reported in this thesis. Further description of the SQL methods specific to each study are reported in chapters 4 to 6.

2.41 Data type

Individual data fields in the BRC Case Register are categorised according to different types of data. The principal data types analysed in the presented studies include:

(i) Varchar: Text containing alphanumeric and punctuation characters up to 8,000 characters in length. These fields may be used to store short pieces of text such as names of diagnoses or medications.

(ii) Text: Text containing alphanumeric and punctuation characters up to 2,147,483,647 characters in length. These fields may be used to store longer pieces of text such as the text from event notes or clinical correspondence.

(iii) Datetime: Calendar date and 24 hour clock time. These fields store the date when a clinical event occurred, e.g. date of hospital admission or date that a clinical document was created.

(iv) Int: Any integer between -2,147,483,647 and 2,147,483,647. These fields store numerical data such as age or number of days spent in hospital.

(v) Null: an empty field containing no data.

2.42 Recoding data

Prior to undertaking statistical analysis, raw data must be recoded into categories that can be meaningfully analysed. In particular, strings of text must be allocated to categories (e.g. diagnostic classification or types of medication) and individual dates or numbers placed into ranges (e.g. age in years). Raw data in the BRC Case Register may be recoded by employing a CASE WHEN statement. CASE WHEN statements have the following components:

(i) CASE: opens the statement

(ii) WHEN: defines a condition

(iii) THEN: defines an output

(iv) ELSE: defines an output to be generated if the data do not match any WHEN conditions

(v) END AS: specifies the name of the recoded variable

CASE WHEN statements with multiple WHEN clauses are processed from the top to the bottom. Once a piece of data has fulfilled a WHEN clause, it will be classified by the associated THEN output and cannot be reclassified by any subsequent WHEN clauses which it might fulfil. The WHEN clause can specify conditions relating to a piece of free text (also known as a 'string', which may include the wildcard '%' to include any text before and/or after the selected string), dates (i.e. exact dates or date ranges), numerical data or null data (i.e. empty field). When multiple clauses are specified in a WHEN statement, they may be combined by using a Boolean operator:

(i) NOT: neither of two conditions are fulfilled

(ii) AND: both conditions are fulfilled

(iii) OR: either condition is fulfilled

The order of precedence for Boolean operators is NOT → AND → OR. This means that in a series of multiple conditions joined by Boolean operators, conditions joined by a NOT will be fulfilled before AND and OR. Brackets are used to override the default order of precedence.

The output from THEN may take the form of a fixed output (e.g. a string of text or a particular number or date) or the output of another variable in the same table. Examples of CASE WHEN statements for recoding clinical data are provided in section 2.51 (diagnosis data) and section 2.52 (medication data).

2.43 Table joins

The ability to join data from different tables into a single output table is the underlying principle of a relational database. Two parameters must be considered when specifying a table join. Firstly, the variable upon which the data will be joined. This is usually an identification variable such as those described earlier in section 2.4. Any data which share the same ID variable upon which the tables are joined may be obtained in a SELECT statement. Further parameters for joining multiple tables include specifying which table serves as the source to which data from another table is joined (i.e. left or right join) and whether data are included if both tables contain matching records (inner join), or if all records from the source table are included (outer join). If the type of join is unspecified, by default an outer join will be performed.

Table joins are specified after the FROM statement in a SQL query. An example is given below:

```
SELECT a.BRCID, a.Gender, b.Ethnicity

FROM Table1 a

LEFT JOIN Table2 b on a.BRCID=b.BRCID
```

In this example, data from Table1 and Table2 are joined by the BRCID. So any variables in rows in which the BRCID match in each table will be selected. The BRCID and Gender are selected from Table1 and Ethnicity is selected from Table2.

The second parameter which must be considered is the underlying unit or “currency” of the data in each table being joined. Each unit of a table is represented by a single row. One row might represent a single patient or it might represent a subunit of a single patient (e.g. one piece of clinical documentation, one hospital admission, one prescribed medication etc). Table joins must be generated such that the currency of all the tables match, otherwise a select statement may duplicate rows of the intended unit. For example, each patient has only one set of demographic parameters at any one time which are usually non-modifiable (e.g. date of birth, ethnicity, gender). In this case, an outer join between a table of clinical data whose rows represent individual patients and another table of demographic data whose rows also represent individual patients would result in a one-to-one mapping of data. However, if each patient has many medication prescriptions (i.e. one-to-many relationship), an outer join involving a table of clinical data whose rows represent individual patients and a table whose rows represent individual medication prescriptions would produce a joined table in the currency of medication prescriptions with duplicate rows of the same patient. In order to avoid errors in duplicating individual rows of a table, it is necessary to ensure that joined tables represent one-to-one relationships of the same underlying currency. Examples of how tables of different currencies may be successfully joined are provided in section 4.2.

2.44 Transforming data: dates

As well as recoding data, it is sometimes necessary to transform numerical data into manageable units for statistical analysis. For example, the month of birth is more usefully analysed as the age of a patient on a particular date. Similarly, the date a patient presented to mental health services and the date they were diagnosed with a particular mental disorder may be more usefully analysed the a time difference in days between the two dates. Date transformations are achieved by using the DATEDIFF command as illustrated in the example below:

```
DATEDIFF(day,[ReferralDate],[DiagnosisDate])
```

In this example, the command produces the delay between the referral date and the diagnosis date in days. Another example (below) illustrates how the age of a patient may be obtained at a particular date:

```
FLOOR((DATEDIFF (day, cleaneddateofbirth, ReferralDate))/365))
```

In this example, the patient's age in years is determined at the referral date by subtracting the date of birth from it and dividing by 365. The FLOOR command then rounds it down to the nearest integer.

2.45 Transforming data: counts

In order to determine the number of times a patient has experienced a particular event (e.g. an admission to hospital), the COUNT command can be used:

```
(SELECT COUNT(*) AS noofadmissions
FROM   sqlcris.dbo.inpatient_episode
WHERE (BrcId = a.brcid) AND
(Admission_Date between DATEADD(m, -12, '1-sep-2014') and '1-sep-2014') AND
(Discharge_Date >= DATEADD(m, -12, '1-sep-2014') OR Discharge_Date = '1 Jan 1900')
GROUP BY BrcId)
FROM Table1 a
```

In this query, for each patient (i.e. BRCID) in Table 1, the number of admissions in the twelve months prior to 1st September 2014 are counted. This is determined by specifying the date range for admissions occurring between twelve months prior to 1st September 2014 (DATEADD(m, -12, '1-sep-2014')) and 1st September 2014 and specifying that the discharge date for admissions must have occurred after the beginning of this date range or the patient must still be in hospital (specified by a discharge date equal to 1st January 1900).

2.5 Types of clinical data obtained

Data for the studies undertaken in this thesis were drawn from the following ePJS-derived SQL tables in the SLaM BRC Case Register:

- EPR: this table contains basic demographic information about the patient.
- Referral: whenever a patient is referred to a SLaM clinical service, an entry is created in the referral table. Each entry includes data on when the referral was made, whether it was accepted and, if so, the date of being accepted and the date of discharge (if still receiving care, the discharge date is set to 1st January 1900).
- Team Episode: a team episode represents a period during which a patient is referred to and receives care from a particular team (if the referral is accepted). A team episode is a unit of a Referral. Therefore, one entry in the Referral table may have many team episodes (e.g. if a patient's care is transferred between multiple teams or they receive care simultaneously from more than one team).
- Inpatient Episode: an inpatient episode is created whenever a patient is admitted to a psychiatric hospital. The principles of this table are analogous to the Referral and Team Episode table.
- MHA Section: the Mental Health Act[114] is a piece of statute legislation in the UK which permits compulsory admission to hospital of people who are thought to be suffering with a serious mental illness of a nature and/or of a degree requiring inpatient treatment and are refusing hospital admission. Whenever a patient is detained under the MHA, the local Mental Health Act office which receives the section papers creates an MHA Section entry which is represented in the MHA Section table.
- Diagnosis: mental disorder diagnosis is recorded in the diagnosis_combined table which is drawn from structured diagnosis fields in ePJS as well as diagnoses recorded

in the unstructured free text which are extracted using natural language processing (described further in section 2.6).

- Medication: details of prescribed medications are stored in the medication_combined table. This includes details of the name of the drug prescribed, the date of prescription, dose and route. Medication data are stored in a composite SQL table which draws data from structured medication fields in ePJS as well as medication data stored in unstructured free text extracted using natural language processing (described further in section 2.6).

Table 2a describe the fields within each of these forms from which data were drawn and their respective data types.

Table 2a: CRIS SQL data structure

CRIS SQL Table Name	CRIS SQL column name	Data field description	Data type	Maximum length (characters)
EPR_Form	cleaneddateofbirth	Month of birth	datetime	
	Gender_ID	Gender	varchar	20
	Marital_Status_ID	Marital Status	varchar	50
	ethnicitycleaned	Ethnicity	varchar	100
	Employment_ID	Employment Status	varchar	50
	Housing_Status	Housing Status	varchar	50
	CN_Doc_ID	Document ID	int	20
	BRCID	Patient ID	int	-
Referral	Referral_Date	Referral Date	datetime	-
	Referral_Status_ID	Referral Status	varchar	50
	Accepted_Date	Accepted Date	datetime	-
	Discharge_Date	Discharge Date	datetime	-
	CN_Doc_ID	Document ID	int	20
	BRCID	Patient ID	int	-
Team_episode	Referral_Date	referral date	datetime	-
	Location_Name	Team name	varchar	150

	Referral_Admin_Status_ID	Status	varchar	35
	Accepted_Date	Accepted date	datetime	-
	Discharge_Date	Discharge date	datetime	-
	CN_Doc_ID	Document ID	int	20
	BRCID	Patient ID	int	-
Inpatient_episode	Admission_Date	Admission Date	datetime	-
	Discharge_Date	Discharge date	datetime	-
	CN_Doc_ID	Document ID	int	20
	BRCID	BRCID	int	-
MHA_Section	MHA_Section_Definition_ID	MHA Section Definition	varchar	100
	Start_Date	Start Date	datetime	-
	CN_Doc_ID	Document ID	int	20
	BRCID	Patient ID	int	-
Medication_combined	Drug	Medication name	varchar	8000
	Dose	Medication dose	varchar	8000
	Start_Date	Start Date	datetime	-
	CN_Doc_ID	Document ID	int	20
	BRCID	Patient ID	int	-
Diagnosis_combined	Primary_Diagnosis	Diagnosis	varchar	8000
	Diagnosis_Date	Diagnosis date	datetime	-
	CN_Doc_ID	Document ID	int	60
	BRCID	Patient ID	int	-

In addition to analysing the data described in Table 2a, unstructured data were also obtained from the event note and correspondence forms in the SLaM BRC Case Register and analysed using the natural language processing methods described subsequently in section 2.6. The event note form is used by clinicians to document individual clinical encounters such as a face-to-face meeting with a patient in an outpatient clinic or on a hospital ward. It is also used to document clinical information outside of face-to-face meetings (e.g. telephone contact or discussion with relatives, third parties and other professionals involved in the patient's care). The correspondence form is used to upload Microsoft Word document files into a patient's record. These files typically include clinical correspondence between healthcare professionals (e.g. a letter to a GP summarising an outpatient assessment) and hospital discharge summaries.

2.51 Diagnosis data

The mental disorder diagnosis recorded by clinicians is an important piece of data which determines clinical management for individual patients and healthcare service delivery. Diagnoses are typically recorded according to the ICD-10 diagnostic classification system.[115] This includes a range of chapters which cover different physiological systems. The chapters of relevance for mental healthcare are Chapter V: Mental and behavioural disorders (F00-F99) and Chapter VI: Diseases of the nervous system (G00-G99). In order to meaningfully analyse these data, it is necessary to recode the various diagnostic codes into larger categories. A SQL script was generated using a series of CASE WHEN statements in order to categorise different diagnostic codes and diagnosis names into larger categories. The script (presented below) was developed and validated by iteratively running and modifying the CASE WHEN statements and manually validating the output when applied to the diagnosis_combined table in order to accurately recode all diagnostic data relating to mental and behavioural disorders in the BRC Case Register. A reduced version of this script was employed in the studies described in this thesis to identify patients with a psychotic disorder. The reduced scripts are presented separately in the supplementary methods sections accompanying the publications incorporated into this thesis.

```
USE [SQLCRIS_User]
select
        [brcid],
        [source_table],
    [diagnosis_date],
    [primary_diagnosis] primary_diagnosisraw,

        case
                when (
                        primary_diagnosis like '%f30%' or
                        primary_diagnosis like '%f31%' or
```

```
                        primary_diagnosis like '%manic%' or
                        primary_diagnosis like '%mania%' or
                        primary_diagnosis like '%bipolar%' or
                        primary_diagnosis like '%bpad%' or
                        primary_diagnosis like '%affective disorder%'
or
                        primary_diagnosis like '%mixed affective%') and
                        primary_diagnosis not like '%trichotillomania%'
and
                        primary_diagnosis not like '%kleptomania%'
                then 'Bipolar'

                when (
                        primary_diagnosis like '%psychotic%' or
                        primary_diagnosis like '%psychosis%' or
                        primary_diagnosis like '%schizophreni%' or
                        primary_diagnosis like '%scizophreni%' or
                        primary_diagnosis like '%schizotyp%' or
                        primary_diagnosis like '%scizotyp%' or
                        primary_diagnosis like '%delusion%' or
                        primary_diagnosis like '%hallucin%' or
                        primary_diagnosis like '%f20%' or
                        primary_diagnosis like '%f21%' or
                        primary_diagnosis like '%f22%' or
                        primary_diagnosis like '%f23%' or
                        primary_diagnosis like '%f24%' or
                        primary_diagnosis like '%f28%' or
                        primary_diagnosis like '%f29%') and
                        primary_diagnosis not like '%affective%' and
                        primary_diagnosis not like '%bipolar%' and
                        primary_diagnosis not like '%bpad%' and
                        primary_diagnosis not like '%mania%' and
                        primary_diagnosis not like '%manic%' and
                        primary_diagnosis not like '%depress%' and
                        primary_diagnosis not like '%mood%' and
                        primary_diagnosis not like '%f3%'
                then 'NonAffectPsychosis'

                when (
                        primary_diagnosis like '%schizoaffective' or
                        primary_diagnosis like '%f25%')
                then 'Schizoaffective'

                when (
                        primary_diagnosis like '%psychosis%' or
                        primary_diagnosis like '%psychotic%' or
                        primary_diagnosis like '%with psyc%') and
                        (primary_diagnosis like '%depress%' or
                        primary_diagnosis like '%f32%' or
                        primary_diagnosis like '%f33%') and
                        primary_diagnosis not like '%without%' and
```

```
                primary_diagnosis not like '%bipolar%'
        then 'PsychoticDepression'

        when (
                primary_diagnosis like '%depress%' or
                primary_diagnosis like '%f32%' or
                primary_diagnosis like '%f33%') and
                (primary_diagnosis not like '%schizophren%'and
                primary_diagnosis not like '%manic%'and
                primary_diagnosis not like '%mania%' and
                primary_diagnosis not like '%bipolar%' and
                primary_diagnosis not like '%f25%' and
                primary_diagnosis not like '%schizoaffective%'
and
                primary_diagnosis not like '%personality%' and
                primary_diagnosis not like '%with psyc%' and
                primary_diagnosis not like '%psychotic
depress%' and
                primary_diagnosis not like '%psychosis and
depression%' and
                primary_diagnosis not like '%psychosis;
depression%')
        then 'UnipolarDepressionWithoutPsychosis'

        when (
                primary_diagnosis like '%cyclothymia%' or
                primary_diagnosis like '%dysthymi%' or
                primary_diagnosis like '%f34%' or
                primary_diagnosis like '%f38%' or
                primary_diagnosis like '%f39%' or
                primary_diagnosis like '%mood disorder%' or
                primary_diagnosis like '%affective disorder%'
or
                primary_diagnosis like '%affective illness%' or
                primary_diagnosis like '%affective symptoms%'
or
                primary_diagnosis like '%mood disturbance%' or
                primary_diagnosis like '%Mood [affective]
disorder%' or
                primary_diagnosis like '%Mood (affective)
disorder%' or
                primary_diagnosis like '%organic mood%')
        then 'OtherAffectiveDisorder'

        when (
                primary_diagnosis like '%psychotic%' or
                primary_diagnosis like '%psychosis%' or
                primary_diagnosis like '%schizophreni%' or
                primary_diagnosis like '%scizophreni%' or
                primary_diagnosis like '%schizotyp%' or
                primary_diagnosis like '%scizotyp%' or
```

```
                primary_diagnosis like '%delusion%' or
                primary_diagnosis like '%hallucin%' or
                primary_diagnosis like '%f20%' or
                primary_diagnosis like '%f21%' or
                primary_diagnosis like '%f22%' or
                primary_diagnosis like '%f23%' or
                primary_diagnosis like '%f24%' or
                primary_diagnosis like '%f28%' or
                primary_diagnosis like '%f29%') and
                (primary_diagnosis like '%affective%' or
                primary_diagnosis like '%bipolar%' or
                primary_diagnosis like '%bpad%' or
                primary_diagnosis like '%mania%' or
                primary_diagnosis like '%manic%' or
                primary_diagnosis like '%depress%' or
                primary_diagnosis like '%mood%' or
                primary_diagnosis like '%f3%')
        then 'OtherAffectivePsychosis'

        when (
                primary_diagnosis like '%f4%' or
                primary_diagnosis like '%anxiety%' or
                primary_diagnosis like '%panic%' or
                primary_diagnosis like '%adjustment%' or
                primary_diagnosis like '%phobic%' or
                primary_diagnosis like '%phobia%' or
                primary_diagnosis like '%ocd%' or
                primary_diagnosis like '%obsessi%' or
                primary_diagnosis like '%compulsi%' or
                primary_diagnosis like '%acute stress%' or
                primary_diagnosis like '%dissociative%' or
                primary_diagnosis like '%somatoform%' or
                primary_diagnosis like '%somatization%' or
                primary_diagnosis like '%bdd%' or
                primary_diagnosis like '%dysmorph%' or
                primary_diagnosis like '%hypochondria%' or
                primary_diagnosis like '%neurosis%' or
                primary_diagnosis like '%neurotic%' or
                primary_diagnosis like '%neurasthenia%' or
                primary_diagnosis like '%fatigue%' or
                primary_diagnosis like '%chronic pain%' or
                primary_diagnosis like '%pain syndrome%' or
                primary_diagnosis like '%pain disorder%' or
                primary_diagnosis like '%briquet%' or
                primary_diagnosis like '%psychosomatic%' or
                primary_diagnosis like '%ptsd%' or
                primary_diagnosis like '%post traumatic%' or
                primary_diagnosis like '%traumatic stress%' or
                primary_diagnosis like '%stress disorder%' or
                primary_diagnosis like '%stress reaction%' or
                primary_diagnosis like '%post-traumatic%' or
```

```
                primary_diagnosis like '%conversion%')
        then 'AnxietyDisorder'

        when (
                primary_diagnosis like '%personality%' or
                primary_diagnosis like '%emotionally unstable%'
or
                primary_diagnosis like '%borderline%' or
                primary_diagnosis like '%histrionic%' or
                primary_diagnosis like '%narcissistic%' or
                primary_diagnosis like '%f60%' or
                primary_diagnosis like '%f61%' or
                primary_diagnosis like '%f62%' or
                primary_diagnosis like '%f69%' and
                primary_diagnosis not like '%learning%')
        then 'PersonalityDisorder'

        when (
                primary_diagnosis like '%f63%' or
                primary_diagnosis like '%f64%' or
                primary_diagnosis like '%f65%' or
                primary_diagnosis like '%f66%' or
                primary_diagnosis like '%f68%' or
                primary_diagnosis like '%identity%' or
                primary_diagnosis like '%gambling%' or
                primary_diagnosis like '%transsex%' or
                primary_diagnosis like '%transvest%' or
                primary_diagnosis like '%fetish%' or
                primary_diagnosis like '%voyeur%' or
                primary_diagnosis like '%paedophil%' or
                primary_diagnosis like '%pedophil%' or
                primary_diagnosis like '%masochis%' or
                primary_diagnosis like '%psychosex%' or
                primary_diagnosis like '%trichotill%' or
                primary_diagnosis like '%kleptoman%' or
                primary_diagnosis like '%pyroman%' or
                primary_diagnosis like '%gender%')
        then 'OtherF6Disorder'

        when (
                primary_diagnosis like '%f00%' or
                primary_diagnosis like '%f01%' or
                primary_diagnosis like '%f02%' or
                primary_diagnosis like '%f03%' or
                primary_diagnosis like '%f04%' or
                primary_diagnosis like '%f05%' or
                primary_diagnosis like '%dementia%' or
                primary_diagnosis like '%cognitive%' or
                primary_diagnosis like '%delirium%' or
                primary_diagnosis like '%alzheimer%' or
                primary_diagnosis like '%lewy%') and
```

```
            primary_diagnosis not like '%F00-F99%'
    then 'DementiaDelirium'

    when (
            primary_diagnosis like '%f06%' or
            primary_diagnosis like '%f07%' or
            primary_diagnosis like '%f09%' or
            primary_diagnosis like '%g0%' or
            primary_diagnosis like '%g1%' or
            primary_diagnosis like '%g2%' or
            primary_diagnosis like '%g3%' or
            primary_diagnosis like '%g4%' or
            primary_diagnosis like '%g8%' or
            primary_diagnosis like '%g9%' or
            primary_diagnosis like '%s06%' or
            primary_diagnosis like '%frontal%' or
            primary_diagnosis like '%intracranial%' or
            primary_diagnosis like '%brain injury%' or
            primary_diagnosis like '%brain damage%' or
            primary_diagnosis like '%brain disease%' or
            primary_diagnosis like '%traumatic brain%' or
            primary_diagnosis like '%head injury%' or
            primary_diagnosis like '%concuss%' or
            primary_diagnosis like '%organic%' or
            primary_diagnosis like '%encaphal%' or
            primary_diagnosis like '%huntingt%' or
            primary_diagnosis like '%parkinson%' or
            primary_diagnosis like '%neuron%' or
            primary_diagnosis like '%sclerosis%' or
            primary_diagnosis like '%epilep%' or
            primary_diagnosis like '%seizure%' or
            primary_diagnosis like '%global amnesia%')
    then 'OtherOrganicBrainDisorder'

    when (
            primary_diagnosis like '%f10%' or
            primary_diagnosis like '%alcohol%')
    then 'AlcoholMisuseDependence'

    when (
            primary_diagnosis like '%f11%' or
            primary_diagnosis like '%f12%' or
            primary_diagnosis like '%f13%' or
            primary_diagnosis like '%f14%' or
            primary_diagnosis like '%f15%' or
            primary_diagnosis like '%f16%' or
            primary_diagnosis like '%f18%' or
            primary_diagnosis like '%f19%' or
            primary_diagnosis like '%cannabis%' or
            primary_diagnosis like '%opioid%' or
            primary_diagnosis like '%opiate%' or
```

```
                        primary_diagnosis like '%cocaine%' or
                        primary_diagnosis like '%amphetamine%' or
                        primary_diagnosis like '%drug%' or
                        primary_diagnosis like '%methadone%')
                then 'DrugMisuseDependence'

                when (
                        primary_diagnosis like '%f50%' or
                        primary_diagnosis like '%f51%' or
                        primary_diagnosis like '%f52%' or
                        primary_diagnosis like '%f53%' or
                        primary_diagnosis like '%f54%' or
                        primary_diagnosis like '%f55%' or
                        primary_diagnosis like '%f59%' or
                        primary_diagnosis like '%eating%' or
                        primary_diagnosis like '%anorexia%' or
                        primary_diagnosis like '%bulimia%' or
                        primary_diagnosis like '%somnia%' or
                        primary_diagnosis like '%sleep%' or
                        primary_diagnosis like '%nightmare%' or
                        primary_diagnosis like '%sexual%' or
                        primary_diagnosis like '%erectile%' or
                        primary_diagnosis like '%ejaculat%' or
                        primary_diagnosis like '%puerper%' or
                        primary_diagnosis like '%postnatal%' or
                        primary_diagnosis like '%postpartum%') and
                        primary_diagnosis not like '%abuse%'
                then 'F5Disorders'

                when (
                        primary_diagnosis like '%f7%' or
                        primary_diagnosis like '%reading%' or
                        primary_diagnosis like '%spelling%' or
                        primary_diagnosis like '%mathematic%' or
                        primary_diagnosis like '%learning%')
                then 'MentalRetardation'

                when (
                        primary_diagnosis like '%f8%' or
                        primary_diagnosis like '%pervasive
development%' or
                        primary_diagnosis like '%pdd%' or
                        primary_diagnosis like '%autis%' or
                        primary_diagnosis like '%asd%' or
                        primary_diagnosis like '%asperger%' or
                        primary_diagnosis like '%language%' or
                        primary_diagnosis like '%speech%' or
                        primary_diagnosis like '%hearing%' or
                        primary_diagnosis like '%development%')
                then 'DevelopmentalDisorder'
```

```
                when (
                        primary_diagnosis like '%f90%' or
                        primary_diagnosis like '%f91%' or
                        primary_diagnosis like '%f92%' or
                        primary_diagnosis like '%f93%' or
                        primary_diagnosis like '%f94%' or
                        primary_diagnosis like '%f95%' or
                        primary_diagnosis like '%f96%' or
                        primary_diagnosis like '%f97%' or
                        primary_diagnosis like '%f98%' or
                        primary_diagnosis like '%adhd%' or
                        primary_diagnosis like '%attention%' or
                        primary_diagnosis like '%hyperactivity%' or
                        primary_diagnosis like '%hyperkinetic%' or
                        primary_diagnosis like '%conduct%' or
                        primary_diagnosis like '%defiant%' or
                        primary_diagnosis like '%encopresis%' or
                        primary_diagnosis like '%enuresis%' or
                        primary_diagnosis like '%tourette%' or
                        primary_diagnosis like '%tic disorder%' or
                        primary_diagnosis like '%tic syndrome%')
                then 'F9Disorder'

                when (
                        primary_diagnosis like '%f99%' or
                        primary_diagnosis like '%fxx%' or
                        primary_diagnosis like 'other' or
                        primary_diagnosis like 'mental illness' or
                        primary_diagnosis like 'mental disorder%' or
                        primary_diagnosis like 'mental state' or
                        primary_diagnosis like '%no Axis 1%' or
                        primary_diagnosis like '%z71.1%' or
                        primary_diagnosis like '%z0%' or
                        primary_diagnosis like '%unspecified disorder%'
or
                        primary_diagnosis like '%unspecified mental%'
or
                        primary_diagnosis like '%xNx%')
                then 'UnrecordedDiagnosis'

                else 'OtherDiagnosis'

        end as primary_diagnosisrecode,

        [cn_doc_id]

  FROM SQLCrisImport.dbo.diagnosis_combined
```

Each of the WHEN statements represents a group of diagnoses (represented in the "primary_diagnosis" column) according to their ICD-10 F codes and strings of text. As the CASE WHEN statements read from top to bottom, as soon as a diagnosis fulfils a WHEN statement, it is categorised according to its associated THEN output. It is possible to specify common stems of diagnoses uses the "%" wildcard in order to capture variations in diagnosis names. For example, '%mania%' captures "mania" as well as "hypomania". However, it would also capture "trichotillomania" and "kleptomania" as these words also contain the common stem "mania". Therefore, it is also necessary to include "not like" statements to filter out these diagnoses from the "Bipolar" category. In this way, each group of diagnoses is classified by using common stems for F codes and names of diagnoses in order to recode the raw data in the "primary_diagnosis" column into a smaller number of categories to facilitate statistical analysis.

2.52 Medication data

Prescription medications are used to treat some patients with physical and mental disorders. Clinicians may prescribe any medication which is licensed within the British National Formulary (BNF)[116] and which are available in local hospital or community pharmacies. Medication data in the BRC Case Register are stored in the medication_combined table. In order to facilitate analysis of medication data in the studies reported in this thesis, it was necessary to recode the raw medication data into categories representing different types of antipsychotic medication by means of CASE WHEN statements. Antipsychotics were categorised according to chapters 4.2.1 and 4.2.2 of the BNF[116] using the SQL script below:

```
USE [SQLCRIS_User]
SELECT [BrcId]
      ,[source_table]
      ,[drug]

      , case
              when (
```

```
                drug like '%benper%' or
                drug like '%anquil%' or
                drug like '%benquil%')
        then 'Benperidol'

        when (
                drug like '%proch%')
        then 'Prochlorperazine'

        when (
                (drug like '%chlorp%' or
                drug like '%chlop%' or
                drug like '%largactil%') and
                drug not like '%chlorphenamine%')
        then 'Chlorpromazine'

        when (
                ((drug like '%depix%' or
                drug like '%depex%' or
                drug like '%depox%' or
                drug like '%flupent%') and
                (drug like '%dec%' or
                drug like '%depot%' or
                drug like '%injection%' or
                drug like '%im%' or
                drug like '%conc%')) and
                drug not like '%tablet%' and
                drug not like '%capsule%' and
                drug not like '%oral%')
        then 'FlupentixolLAI'

        when (
                drug like '%depix%' or
                drug like '%depex%' or
                drug like '%depox%' or
                drug like '%flupent%' or
                drug like '%fluan%')
        then 'Flupentixol'

        when (
                ((drug like '%haloper%' or
                drug like '%haliper%' or
                drug like '%dozic%' or
                drug like '%hald%' or
                drug like '%hadol%' or
                drug like '%seren%') and
                (drug like '%dec%' or
                drug like '%depot%')) and
                drug not like '%tablet%' and
                drug not like '%capsule%' and
                drug not like '%oral%')
```

```
                then 'HaloperidolLAI'

                when (
                        drug like '%haloper%' or
                        drug like '%haliper%' or
                        drug like '%dozic%' or
                        drug like '%hald%' or
                        drug like '%hadol%' or
                        drug like '%seren%')
                then 'Haloperidol'

                when (
                        (drug like '%levom%' or
                        drug like '%methotrim%' or
                        drug like '%nozinan%') and
                        drug not like '%levomenthol%')
                then 'Levomepromazine'

                when (
                        (drug like '%pericy%' or
                        drug like '%perici%' or
                        drug like '%neula%') and
                        drug not like '%neulasta%')
                then 'Pericyazine'

                when (
                        drug like '%perph%' or
                        drug like '%fentaz%')
                then 'Perphenazine'

                when (
                        drug like '%pimoz%' or
                        drug like 'orap%')
                then 'Pimozide'

                when (
                        drug like 'promaz%')
                then 'Promazine'

                when (
                        (drug like '%amis%' or
                        drug like '%ahilsul%' or
                        drug like '%amilsul%' or
                        drug like '%amosul%' or
                        drug like '%amsulp%' or
                        drug like '%amylsus%' or
                        drug like '%asulpri%' or
                        drug like '%solian%') and
                        drug not like '%lamisil%')
                then 'Amisulpiride'
```

```
                when (
                        (drug like '%sulpi%' or
                        drug like '%sulpr%' or
                        drug like '%sulpa%' or
                        drug like '%sulpe%' or
                        drug like '%sulpo%' or
                        drug like '%sulpu%' or
                        drug like '%dolm%') and
                        drug not like '%ami%')
                then 'Sulpiride'

                when (
                        drug like '%trifluperidol%' or
                        drug like '%triper%')
                then 'Trifluperidol'

                when (
                        drug like '%trifl%' or
                        drug like '%stelaz%')
                then 'Trifluoperazine'

                when (
                        ((drug like '%zucl%' or
                        drug like '%clopixol%') and
                        (drug like '%depot%' or
                        drug like '%dec%')) and
                        drug not like '%acetate%' and
                        drug not like '%acuph%' and
                        drug not like '%tablet%' and
                        drug not like '%capsule%' and
                        drug not like '%oral%')
                then 'ZuclopenthixolLAI'

                when (
                        drug like '%zucl%' or
                        drug like '%clopixol%' or
                        drug like '%acuph%')
                then 'Zuclopenthixol'

                when (
                        ((drug like '%maintena%' and
                        drug not like '%tablet%' and
                        drug not like '%capsule%' and
                        drug not like '%oral%') or
                        ((drug like '%arip%' or
                        drug like '%arpip%' or
                        drug like 'abil%') and
                        (drug like '%depot%' or
                        drug like '%400%'))) and
                        drug not like '%tablet%' and
                        drug not like '%capsule%' and
```

```
                drug not like '%oral%' and
                drug not like '%maintenance%')
        then 'AripiprazoleLAI'

        when (
                drug like '%arip%' or
                drug like '%arpip%' or
                drug like 'abil%')
        then 'Aripiprazole'

        when (
                drug like '%cloz%' or
                drug like '%cloxap%' or
                drug like '%denz%' or
                drug like '%zapo%')
        then 'Clozapine'

        when (
                (((drug like '%olanz%' or
                drug like '%olaza%') and
                (drug like '%depot%' or
                drug like '%embon%')) and
                drug not like '%tablet%' and
                drug not like '%capsule%' and
                drug not like '%oral%') or
                (drug like '%zypa%' and
                drug not like '%tablet%' and
                drug not like '%capsule%' and
                drug not like '%oral%'))
        then 'OlanzapineLAI'

        when (
                drug like '%olanz%' or
                drug like '%olaza%' or
                drug like '%zypr%')
        then 'Olanzapine'

        when (
                ((drug like '%palip%' and
                (drug like '%depot%' or
                drug like '%palmi%' or
                drug like '%injection%')) and
                drug not like '%tablet%' and
                drug not like '%capsule%' and
                drug not like '%oral%') or
                (drug like '%xeplion%' and
                drug not like '%tablet%' and
                drug not like '%capsule%' and
                drug not like '%oral%'))
        then 'PaliperidoneLAI'
```

```
                when (
                        drug like '%palip%' or
                        drug like '%inveg%')
                then 'Paliperidone'

                when (
                        drug like '%quet%' or
                        drug like '%seroq%')
                then 'Quetiapine'

                when (
                        ((drug like '%consta%' and
                        drug not like '%tablet%' and
                        drug not like '%capsule%' and
                        drug not like '%oral%') or
                        (drug like '%risp%' and
                        (drug like '%depot%' or
                        drug like '%injection%'))) and
                        drug not like '%tablet%' and
                        drug not like '%capsule%' and
                        drug not like '%oral%' and
                        drug not like '%crispbread%')
                then 'RisperidoneLAI'

                when (
                        drug like '%risp%' and
                        drug not like '%crispbread%')
                then 'Risperidone'

                when (
                        ((drug like '%fluph%' and
                        (drug like '%depot%' or
                        drug like '%dec%' or
                        drug like '%injection%')) and
                        drug not like '%tablet%' and
                        drug not like '%capsule%' and
                        drug not like '%oral%') or
                        (drug like '%modec%' and
                        drug not like '%tablet%' and
                        drug not like '%capsule%' and
                        drug not like '%oral%'))
                then 'FluphenazineLAI'

                when (
                        drug like '%fluph%' or
                        drug like '%modit%')
                then 'Fluphenazine'

                when (
                        drug like '%pipo%')
                then 'PipotiazineLAI'
```

```
when (
        drug like '%asen%' or
        drug like '%sycr%')
then 'Asenapine'

when (
        drug like '%luras%' or
        drug like '%latud%')
then 'Lurasidone'

when (
        drug like '%zotep%' or
        drug like '%zolep%')
then 'Zotepine'

when (
        drug like '%sertin%' or
        drug like '%serdol%')
then 'Sertindole'

when (
        drug like '%thior%' or
        drug like '%meller%')
then 'Thioridazine'

when (
        drug like 'loxa%')
then 'Loxapine'

when (
        drug like '%oxyper%')
then 'Oxypertin'

when (
        drug like '%drope%' or
        drug like '%drolep%')
then 'Droperidol'

when (
        drug like '%flus%' or
        drug like '%redep%')
then 'Fluspirilene'

when (
        drug like '%remox%' or
        drug like '%roxiam%')
then 'Remoxipride'

else 'OtherDrug'
```

```
        end as drugrecoded

      ,[drug_type]
      ,[status]
      ,[tense]
      ,[start_date]
      ,[End_Date]
      ,[dose]
      ,[ViewDate]
      ,[ViewText]
      ,[CN_Doc_ID]
      ,[dose_value]
      ,[dose_unit]
      ,[frequency]
      ,[time_unit]
      ,[interval]
      ,[when]
      ,[route]
      ,[id]
  FROM [SQLCrisImport].[dbo].[medication_combined]
```

Many drugs have both a generic name which defines the active compound in the preparation as well as a registered trade name which is the name given to the preparation by the pharmaceutical company which produces it. For each antipsychotic recorded in the BNF, a WHEN statement was generated in order to capture the minimum word stem to define the generic and trade names of each drug. Further statements were created in order to distinguish between antipsychotic drugs administered as long acting injectable depot preparations (suffixed "LAI" in the THEN output) and oral or short acting injectable preparations. The script was validated by manually reviewing its output when applied to the medication_combined table to ensure accuracy of classification of the CASE WHEN statements. Once validated, the script was stored as a view ("[SQLCRIS_User].[RPatel].[rp_medication_combined_antipsychotic_recode]") from which antipsychotic prescription data was extracted in the studies presented in this thesis. In order to allow for data on discontinued antipsychotics to be extracted, previous editions of the BNF were reviewed from September 1992 up to October 2013 to identify the names of discontinued antipsychotics. These data are presented in Table 2b and 3b.

Table 2b: Data from the 66th edition (October 2013) of the British National Formulary (BNF) used to recode antipsychotic medication data

Generic drug name	**Trade name**	**Drug class**	**Route**	**BNF recommended dose range**	**BNF maximum per day (if stated)**	**BNF Edition added**	**Other notes**
Benperidol	Anquil/Benquil	FGA	Oral	0.25mg-1.5mg	1.5mg		
Chlorpromazine Hydrochloride	Largactil	FGA	Oral/IM/PR	25mg-300mg	1000mg		
Flupentixol/Flupenthixol	Depixol/Fluanxol	FGA	Oral	3mg-18mg	18mg		
Haloperidol	Dozic/Haldol/Serenace	FGA	Oral/IM	0.5mg-15mg	30mg		
Levomepromazine	Nozinan	FGA	Oral	25mg-1000mg	1000mg		Name changed from "Methotrimeprazine" in BNF 43 (Mar02)
Pericyazine/Periciazine	Neulactil	FGA	Oral	25mg-300mg	300mg		
Perphenazine	Fentazin	FGA	Oral	4mg-24mg	24mg		
Pimozide	Orap	FGA	Oral	2mg-20mg	20mg		
Prochlorperazine	(Buccastem/Stemetil)	FGA	Oral/IM	12.5mg-100mg	100mg		
Promazine Hydrochloride		FGA	Oral	25mg-800mg	800mg		
Sulpiride	Dolmatil/Sulpor/Sulpitil/Sulparex	FGA	Oral	200mg-2400mg	2400mg		
Trifluoperazine	Stelazine	FGA	Oral	1mg-10mg (no max stated)			
Zuclopenthixol	Clopixol	FGA	Oral	2mg-50mg	150mg		Name changed from "Zuclopenthixol Dihydrochloride" in BNF 55 (Mar08)
Zuclopenthixol Acetate	Clopixol/Acuphase	FGA	IM	50mg-150mg	150mg		

Amisulpride	Solian	SGA	Oral	50mg-800mg	1200mg	BNF 35 (Mar98)	
Aripiprazole	Abilify	SGA	Oral/IM	5mg-15mg	30mg	BNF 48 (Sep04)	
Clozapine	Clozaril/Denzapine/Zaponex	Clozapine	Oral	12.5mg-900mg	900mg		
Olanzapine	Zyprexa	SGA	Oral	2.5mg-20mg	20mg	BNF 33 (Mar97)	
Paliperidone	Invega	SGA	Oral	3mg-12mg	12mg	BNF 55 (Mar08)	
Quetiapine	Seroquel	SGA	Oral	25mg-800mg	800mg	BNF 35 (Mar98)	
Risperidone	Risperdal	SGA	Oral	0.5mg-16mg	16mg	BNF 26 (Sep93)	
Flupentixol Decanoate	Depixol	FGA	LAI	20mg-400mg	400mg		
Fluphenazine Decanoate	Modecate	FGA	LAI	12.5mg-100mg	100mg		
Haloperidol	Haldol Decanoate	FGA	LAI	12.5mg-300mg	300mg		
Olanzapine Embonate	ZypAdhera	SGA	LAI	210mg-405mg	405mg	BNF 59 (Mar10)	
Paliperidone	Xeplion	SGA	LAI	25mg-150mg	150mg	BNF 62 (Sep11)	
Pipotiazine Palmitate	Piportil Depot	FGA	LAI	25mg-200mg	200mg		
Risperidone	Risperdal Consta	SGA	LAI	25mg-50mg	50mg	BNF 45 (Mar03)	
Zuclopenthixol Decanoate	Clopixol	FGA	LAI	100mg-600mg	600mg		
Asenapine	Sycrest	SGA	Oral	5mg-20mg	20mg	BNF 63 (Mar12)	

FGA: First generation antipsychotic SGA: Second generation antipsychotic IM: Intramuscular PR: Per rectum LAI: Long acting injectable Maximum dose range for oral preparations: per day Maximum dose range for LAI preparations: per dose

Table 2c: Historical data from previous editions of the British National Formulary (BNF) used to recode antipsychotic medication data

Generic drug name	Trade name	Drug class	Route	BNF recommended dose range	BNF maximum per day (if stated)	BNF Edition added	BNF Edition removed
Zotepine	Zoleptil	SGA	Oral	25mg-300mg	300mg	BNF 37 (Mar99)	BNF 61 (Mar11)
Sertindole	Serdolect	SGA	Oral	4mg-20mg	24mg	BNF 33 (Mar97)	BNF 60 (Sep10)
Fluphenazine Hydrochloride	Moditen	FGA	Oral	1mg-20mg	20mg		BNF 55 (Mar08)
Thioridazine	Melleril	FGA	Oral	10mg-300mg	600mg		BNF 53 (Mar07)
Loxapine	Loxapac	FGA	Oral	10mg-100mg	250mg		BNF 45 (Mar03)
Oxypertin		FGA	Oral	10mg-300mg	300mg		BNF 43 (Mar02)
Droperidol	Droleptan	FGA	Oral/IM	5mg-120mg	120mg		BNF 41 (Mar01)
Fluspirilene	Redeptin	FGA	LAI	2mg-8mg	20mg		BNF 31 (Mar96)
Trifluperidol	Triperidol	FGA	Oral	0.5mg-8mg	8mg		BNF 29 (Mar95)
Remoxipride Hydrochloride	Roxiam	FGA	Oral	150mg-450mg	600mg		BNF 28 (Sep94)
Fluphenazine Enanthate	Moditen Enanthate	FGA	LAI	12.5mg-100mg	100mg		BNF 24 (Sep92)

FGA: First generation antipsychotic
SGA: Second generation antipsychotic
IM: Intramuscular
LAI: Long acting injectable
Maximum dose range for oral preparations: per day
Maximum dose range for LAI preparations: per dose

2.6 Natural Language Processing

NLP methods developed using the General Architecture for Text Engineering software package (GATE) were employed to extract clinical data from unstructured free text event notes and correspondence documents in the BRC Case Register.[99] NLP applications may be developed by employing a rules-based approach or a machine learning approach (or a hybrid of both methods). Prior to applying either approach, the body of text to be analysed (also known as a "corpus") is pre-processed:

(i) Tokenisation: the corpus of text is classified into individual tokens. These may represent a single word, a space between words or a punctuation mark.

(ii) Sentence splitting: sentences within the corpus are delineated by identifying tokens that represent breaks between sentences. By default, these are full stops (i.e. ".") but can also include line-breaks and other punctuation marks.

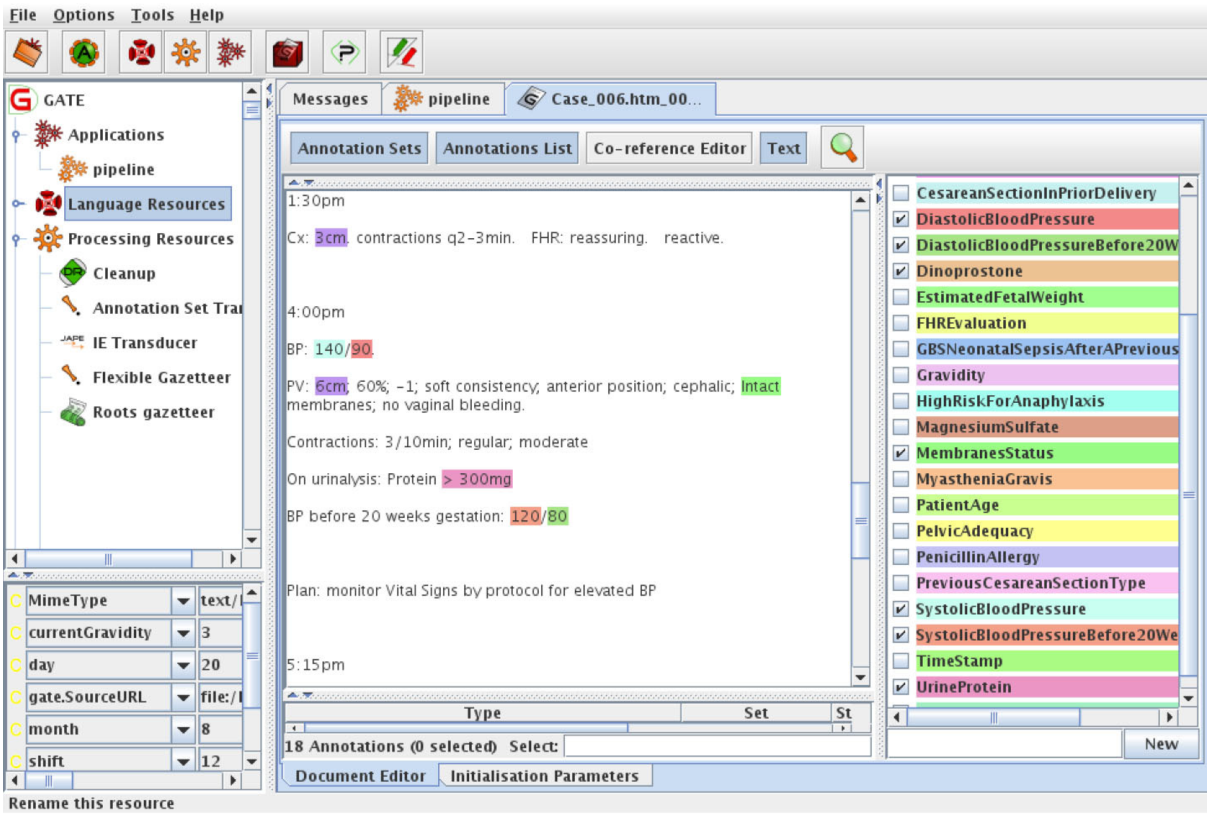
(iii) Parts of speech tagging: word tokens are classified as parts of speech (e.g. noun, verb, adjective etc)

(iv) Gazetteer classification: each word token is classified by a pre-specified dictionary of terms known as a gazetteer. The gazetteer is developed by selecting key words of relevance to the application being developed. For example, an NLP application to determine someone's place of birth may include a gazetteer containing names of different countries or cities.

2.61 Rules-based NLP applications

In a rules-based approach, pre-specified rules are applied to individual sentences within a body of text (known as a corpus) to extract information. The rules typically include a "trigger" which flags a sentence as potentially containing useful information. A series of rules are then applied to the sentence in order to extract and summarise the information of interest. Figure 2a illustrates the output of a GATE NLP application designed to extract useful information from obstetric clinical records using a rules-based approach.

Figure 2a: GATE NLP application to extract clinical information from obstetric notes developed by the Department of Computer Science, University of Sheffield[99]



In this example, rules are specified to extract information on degree of cervical dilatation (highlighted in purple), current systolic and diastolic blood pressure (highlighted cyan and red), systolic and diastolic blood pressure before 20 weeks gestation (highlighted red and green) and membrane status (highlighted light green).

2.62 Machine learning NLP applications

An alternative approach to developing NLP applications is to use a support vector machine learning (SVM) approach.[100] SVM is a mathematical technique in which multidimensional data are automatically classified based on an algorithm derived from prior analysis of an independent 'training' dataset. Data of multiple dimensions are transformed by a linear classifier which separates individual data points by pre-specified classes (or features). SVM is widely used in the analysis of neuroimaging data, especially in psychosis,[117] but has only recently been applied to EHR data.

NLP applications may be developed in GATE using SVM methods to automatically classify unstructured text data. After tokenisation and sentence splitting, a gazetteer of key words is specified which includes words that are relevant to the construct of interest (e.g. “cannabis”, “hash”, “skunk”, “resin”, “marijuana” for an application to identify cannabis use). A free text search is performed to identify any sentence containing a key word of interest. A human annotator then reads through a sample of sentences and manually classifies them according to whether they or not they contain the construct of interest. The human-classified sentences form the ‘training’ dataset upon which an SVM classifier is derived by analysing how tokens cluster around the key word of interest in each sentence (known as a “bag-of-words” approach) in order automatically separate between the presence or absence of the construct. The optimum classifier is determined by a cross-validation process in which the training data set is split into ten groups. Nine groups are used to train the classifier with the last used to test its accuracy. This process is repeated 10 times in order to develop the most accurate classifier. The resulting classifier is then tested on an external dataset (described further in section 2.63) to evaluate its accuracy independent of the training dataset, before being applied to a corpus of data to extract information of interest.

More complex applications may include more than just two classes (e.g. “good”, “intermediate” or “poor” eye contact), or an additional “unknown” class for sentences that do not explicitly infer the presence or absence of a particular construct. A hybrid application may also be developed in which an SVM-derived NLP application is first run on a corpus of text to perform an initial filter before a more specific rules-based approach is applied.

2.63 Precision testing

It is necessary to evaluate the performance of an NLP application with precision testing in order to determine how accurately it extracts and classifies information. In order to do this, a reference dataset of sentences are classified by at least one human annotator. The newly developed NLP application is then applied to this reference dataset and the output from the NLP application compared with the human annotator classification to determine the NLP accuracy in terms of its precision (positive predictive value) and recall (sensitivity). These are defined in Figure 2b, below.

Figure 2b: Precision statistics used to evaluate the performance of an NLP application

		Human Annotator		
		Positive	Negative	
NLP outcome	Positive	True positive	False positive	**Precision** Positive predictive value = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ NLP positive}}$
	Negative	False negative	True negative	Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ NLP negative}}$
		Recall Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Annotator positive}}$	Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Annotator negative}}$	

Precision refers to the proportion of true positive sentences classified by the NLP application out of the total number of sentences classified as positive by the NLP application. Recall refers the proportion of true positive sentences classified by the NLP application out of the total number of sentences classified as positive by the human annotator. Thus, an application with a high degree of precision is necessary to reduce the frequency of false positive classifications and a high degree of recall is necessary to reduce the frequency of false negative classifications.

2.64 Active learning

Active learning is an extension of the SVM NLP method described in section 2.62. The technique involves first deriving and applying an SVM NLP application over a text corpus. The automated SVM algorithm aims to separate each data point (in this case, each sentence) into different classes by transforming the data using a linear classifier, also known as a "hyperplane". When the linearly transformed data are plotted, the hyperplane represents the line which best separates the data into the different classes. Some data points (which, in the case of NLP, represent individual sentences) may be further away from the hyperplane than others. The further away a data point is from the hyperplane, the better the SVM classifier is able to distinguish which class it belongs to. The distance from the hyperplane (known as the "margin") is represented as a proportion of the maximum possible distance from the hyperplane between -1 and +1. Values between 0 and +1 represent data points which are correctly classified while values between -1 and 0 represent data points which are incorrectly classified.

In active learning, sentences with a low margin (which are difficult for the SVM algorithm to classify) are selected and presented to a human annotator to manually classify. These manually classified sentences then form the basis of additional training data which are re-analysed in an attempt to develop a more accurate SVM classifier. This process is known as active learning and may be repeated one or more times to improve the performance of an SVM-derived NLP application.

2.65 Setting a minimum margin threshold to increase precision

It is possible to further customise the parameters of an SVM-derived NLP application by adjusting the minimum threshold for the margin (i.e. the distance of a data point from the hyperplane) to accept the output of the NLP application. In this way, the NLP application will only offer an output on sentences which are classified with a high degree of confidence. This has the effect of increasing the precision of the NLP application (increasing the proportion of true positive sentences classified), but at the expense of reducing the recall (increasing the proportion of false negative sentences classified). Different thresholds of SVM margin result in different precision and recall statistics. The available range of performance may be plotted on a Receiver Operator Characteristic (ROC) curve in order to select the optimum precision and recall parameters for the purposes of the NLP application. An example of this technique in relation to developing an NLP application to ascertain cannabis use is described further in chapter 4.

2.66 TextHunter

The SVM NLP techniques described in 2.62 to 2.64 were performed using the TextHunter software package.[118] TextHunter is a bespoke frontend graphical user interface (GUI) for the GATE software package. The software facilitates each of the steps involved in developing an SVM NLP application from gazetteer development, sentence searching, human annotation, SVM NLP development, SVM NLP precision testing and application of the resulting NLP application to a corpus of text. TextHunter is based on Java architecture and may be run on any Windows PC which has a Java Runtime Environment and GATE installed. TextHunter is specifically designed to interface directly with data in the BRC Case Register using SQL. A detailed description of TextHunter is provided in chapter 6 of this thesis as a conference article.[119]

In summary, a dictionary of key words is specified using regular expressions. TextHunter then runs an initial search phase in which sentences in event notes and/or correspondence in the BRC Case Register containing the key words of interest are extracted and stored as a SQL table whose currency is each individual sentence containing one (or more) of the key words specified. The next phase involves human annotation of a reference standard dataset and a training dataset. The reference standard dataset (known as the “gold” dataset) typically contains around 300 sentences and should also be annotated independently by a second human annotator. Inter-annotator agreement (IAA) is estimated between the two human annotators as the percentage of observed agreement and

Cohen's kappa. If a reasonable degree of IAA is established, the first annotator proceeds to annotate a training dataset (known as the "seed" dataset) of around 500 sentences. At this stage, TextHunter runs the machine learning module in GATE using the training "seed" dataset over 10 folds of cross validation to develop the optimum SVM algorithm (with the highest "F1" value, which represents the harmonic mean of precision and recall). The final stage is to apply the optimum model developed during analysis of the training dataset upon the "gold" reference standard dataset to estimate precision statistics. Further rounds of active learning annotation and SVM model development may be performed to improve precision and recall. The resulting NLP application is then applied by TextHunter to event and/or correspondence data in the BRC Case Register SQL database and stored in a separate table. The data may then be accessed by performing a table join on the NLP data using the methods previously described in section 2.43.

2.67 NLP applications employed in this thesis

All the studies presented in this thesis employed two previously-established rules-based NLP applications to extract diagnosis and medication data from the BRC Case Register. These applications have been used in a number of previously published studies to extract data on diagnosis and prescribed medications at individual patient level. Further novel SVM NLP applications were specifically developed for this thesis to extract data to investigate cannabis use in first episode psychosis (chapter 4) and negative symptoms in schizophrenia (chapter 5). Details on their development are provided in the supplementary methods section of the respective chapters.

2.7 Preparing BRC Case Register data for statistical analysis

Statistical analysis for all studies reported in this thesis was performed using STATA software version 12.[120] The specific statistical methods employed in each study are described in their respective chapters (4, 5 and 6). This section describes how BRC Case Register data were prepared for statistical analysis. Data were recoded into categorical variables (e.g. gender, ethnicity, marital status etc) using the methods described in section 2.42. After recoding variables, the data for analysis were exported from the SQL database into Microsoft Excel 2010 spread sheets. The first row of each sheet to be analysed specified the variable names for each column of data. In Microsoft Excel, the data type for each column of data was specified (i.e. numerical, date, text etc). The spread sheet was then imported into STATA (again, using the first row of the spread sheet to specify variable names). Each STATA dataset was analysed by creating “do” files to run a series of statistical commands upon the dataset to produce the statistical output for each analysis.

3. Clinical outcomes in people with first episode psychosis who present to high-risk clinical services

This chapter incorporates an article published in Acta Psychiatrica Scandinavica based on the use of CRIS to investigate clinical outcomes in people with first episode psychosis who had initially presented to high-risk clinical services. The article and accompanying supplementary data are presented in section 3.1. Further supplementary methods describing SQL data extraction and statistical analysis using STATA are described in section 3.2.

3.1 Acta Psychiatrica Scandinavica journal article

Please see overleaf.

Acta Psychiatrica Scandinavica

Acta Psychiatr Scand 2015: 1–10

DOI: 10.1111/acps.12480

ACTA PSYCHIATRICA SCANDINAVICA

Services for people at high risk improve outcomes in patients with first episode psychosis

Fusar-Poli P, Díaz-Caneja CM, Patel R, Valmaggia L, Byrne M, Garety P, Shetty H, Broadbent M, Stewart R, McGuire P. Services for people at high risk improve outcomes in patients with first episode psychosis.

Objective: About one-third of patients referred to services for people at high risk for psychosis may have already developed a first episode of psychosis (FEP). We compared clinical outcomes in FEP patients who presented to either high risk or conventional mental health services.
Method: Retrospective study comparing duration of hospital admission, referral-to-diagnosis time, need for compulsory hospital admission and frequency of admission in patients with FEP who initially presented to a high-risk service (n = 164) to patients with FEP who initially presented to conventional mental health services (n = 2779). Regression models were performed, controlling for several confounders.
Results: FEP patients who had presented to a high-risk service spent 17 fewer days in hospital [95% CI: −33.7 to (−0.3)], had a shorter referral-to-diagnosis time [B coefficient −74.5 days, 95% CI: −101.9 to −(47.1)], a lower frequency of admission [IRR: 0.49 (95% CI: 0.39–0.61)] and a lower likelihood of compulsory admission [OR: 0.52 (95% CI: 0.34–0.81)] in the 24 months following referral, as compared to FEP patients who were first diagnosed at conventional services.
Conclusion: Services for people at high risk for psychosis are associated with better clinical outcomes in patients who are already psychotic.

P. Fusar-Poli[1,2,3,a], C. M. Díaz-Caneja[1,4,a], R. Patel[1,a], L. Valmaggia[2,5], M. Byrne[1,2], P. Garety[5], H. Shetty[6], M. Broadbent[6], R. Stewart[7], P. McGuire[1,2]

[1]King's College London, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, London, [2]South London and Maudsley NHS Foundation Trust, OASIS, London, [3]Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy, [4]School of Medicine, Child and Adolescent Psychiatry Department, Hospital General Universitario Gregorio Marañón, IiSGM, CIBERSAM, Universidad Complutense, Madrid, Spain, [5]King's College London, Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, London, , [6]South London and Maudsley NHS Foundation Trust, Biomedical Research Centre Nucleus, London, and [7]King's College London, Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, London, UK

Key words: first episode psychosis; psychosis risk; UHR; ARMS; schizophrenia; prodromal; CRIS; SLaM

Paolo Fusar-Poli, Department of Psychosis Studies, Institute of Psychiatry PO63, De Crespigny Park, SE5 8AF London, UK. E-mail: paolo.fusar-poli@kcl.ac.uk

[a]Equal contributions to authorship

Accepted for publication August 4, 2015

Significant outcomes

- First episode psychosis (FEP) patients who present to high-risk services have a lower frequency of hospital admission, fewer days in hospital and a lower likelihood of compulsory admission than FEP patients who present to conventional services.
- These better outcomes may reflect the engagement of patients at an earlier stage of the FEP.
- Services for people at high-risk may benefit patients who are already psychotic by facilitating earlier detection.

Limitations

- Observational study: patients were not randomly assigned to the samples that were compared.
- It was not possible to control for all treatments received from the FEP diagnosis over the follow-up time.
- However, the use of antipsychotic exposure as proxy index of illness severity and treatment offered after the diagnosis of FEP did not affect the findings.

Introduction

Over the last two decades, specialised clinical services have been developed for people at high risk for psychosis (1, 2). Providing clinical care at this stage may reduce the risk of these individuals subsequently developing a psychotic disorder (3, 4). However, although they are designed for people who are vulnerable to psychosis, about a third of those referred are found to already be in the first episode of psychosis (FEP) when they are assessed by specialised high-risk teams (1, 5). This may reflect the fact that the symptoms associated with the high-risk state are qualitatively similar to those of first episode psychosis, but less severe. As soon as a diagnosis of psychosis is confirmed, high-risk services immediately refer these patients to specialised first episode services, where specific treatment can be initiated.

In the absence of a specialised service for people at high risk, an individual with signs of high-risk state would not usually be referred for mental health care: a referral would only be made if the patient was considered to be psychotic. Consequently, a person who was psychotic but incorrectly perceived to be at high risk is likely to be assessed by mental health services sooner when there is a high-risk team available. Previous studies have highlighted that the longer the delay between the onset of psychosis and the initiation of treatment, the poorer the outcome (6). This suggests that first episode patients who are inadvertently referred to high-risk services have better clinical outcomes than those whose first contact is with conventional mental health services.

Aims of the study

This study tested the hypothesis that first episode patients who presented to a high-risk service had better clinical outcomes compared to those presenting to conventional mental health services, as indexed by number and duration of hospital admissions, referral-to-diagnosis time and the proportion of admissions that were compulsory.

Material and methods

Data sources

A retrospective review of clinical records was performed using data from the South London and Maudsley NHS Trust (SLaM) electronic Patient Journey System (ePJS), including all information documented by professionals involved in each patient's clinical care. ePJS anonymised clinical data from over 250 000 patients receiving secondary mental health care are available in the SLaM Biomedical Research Council (BRC) Case Register, which facilitates focussed searching and data extraction from structured and unstructured text fields within the electronic health record using the Clinical Record Interactive Search tool (CRIS) (7).

Samples

First episode patients referred to a high-risk service. These patients were drawn from referrals to OASIS, a clinical service for people at high risk for psychosis in South London (1). On average, approximately one in three of those referred to OASIS meet criteria for a FEP (1). Patients with this diagnosis are assertively referred to an affiliated clinical team specialised in the management of first episode psychosis. OASIS was first implemented in the boroughs of Lambeth and Southwark before later being implemented in the boroughs of Lewisham and Croydon. From January 2001 to September 2011, there were 1090 referrals to OASIS. For data prior to 2007, electronic records of clinical files were not complete, but a scanned copy of the written files was available for manual review.

Between OASIS inception and September 2011, 263 referrals received a diagnosis of psychosis. Of those, 34 received a diagnosis of multiple episode psychosis, whereas two patients had been referred on two occasions to the high-risk service and were duplicated in the referral log 10 patients initially diagnosed with an at risk state for psychosis (at risk mental state, ARMS) and subsequently reported to have made the transition within 3 months were also included as first episode psychosis, because a retrospective re-assessment of their symptoms after they were diagnosed with psychosis made the clinicians consider that they were already psychotic when they were referred to the service (5, 8). Thus, the initial first episode sample consisted of 237 patients. Of those, nine cases were not available in the ePJS. After careful clinical file review, 27 patients did not fulfil criteria for a FEP: three patients were actually diagnosed with an at risk mental state and did not make the transition; nine were diagnosed with multiple episode psychosis, and fifteen did not receive a primary diagnosis of psychosis (six patients were diagnosed with affective disorders without psychotic symptoms, four with personality disorders, two with adjustment disorders, one with substance use disorders, one with attention deficit hyperactivity disorder and one with a learning disability). Thus, our sample comprised 201 patients with a confirmed diagnosis of first episode psychosis. Of those, in 37 cases, there was no assessment or assertive intervention by OASIS (four patients did not engage and there were not enough signs of concern to assertively refer them to other specialised services, whereas 33 patients had already passed the psychosis threshold and were screened out before the assessment took place). Thus, the final sample comprised 164 cases with a confirmed diagnosis of first episode psychosis and whose initial management included an active intervention from OASIS (assessment and/or assertive referral to more appropriate services) (see Figure S1).

First episode patients who presented to conventional mental health services. In the catchment area served by SLaM, patients thought or identified to have first episode psychosis usually present to generic adult mental health community, home treatment and inpatient teams, or directly to specialised first episode services. The generic mental health teams may subsequently refer these patients to the first episode services, which in SLaM include the Lambeth Early Onset service (LEO), Southwark Team for Early Intervention in Psychosis (STEP), Lewisham Early Intervention Service (Lewisham EIS) and Croydon Outreach Assessment Support Team (COAST). Like OASIS, these first episode services accept self-referrals and referrals made by health and non-health agencies (9) and provide a similar form of clinical care which focuses on assertive patient engagement and early clinical intervention (1). In this study, we controlled for potential differences in the provision of treatment by including borough of residence and antipsychotic exposure as covariates in all multivariable analyses.

We compared clinical outcomes of patients whose first contact with mental health services for first episode psychosis was either with a specialised high-risk service or with conventional services. The conventional services sample (n = 2779) was drawn from all patients who presented for the first time between 2007 and 2011 and received a diagnosis of first episode psychosis. The period of 2007–2011 was chosen as 2007 was the first full year in which electronic records were used across all SLaM services. The sample was filtered to exclude any patients who had previously been referred to OASIS with first episode psychosis (in order to ensure that individuals analysed in the high-risk group were not duplicated in the conventional services group) and to only include patients aged between 14 and 35 at the time of referral to SLaM (to reflect the inclusion criteria of the high-risk service). Most (about 80%) of this sample (n = 2284) presented to generic adult mental health services; a minority presented directly to first episode teams (n = 495). Clinical record data for the entire sample were accessed using CRIS, a bespoke software designed to rapidly search electronic records (7).

Data collection

The following data were extracted from both samples: sociodemographic characteristics (age, gender, ethnicity, marital and employment status, borough of residence), diagnosis, referral-to-diagnosis time (measured in days from the date of referral to date of recording of diagnosis), date of first antipsychotic prescription, dates of hospital admissions in a 24-month follow-up period, dates of compulsory admissions under the UK Mental Health Act [MHA; which regulates involuntary admission to hospital of people diagnosed with a mental disorder for assessment and/or treatment (10)] in a 24-month follow-up period and the total cumulative duration of hospital admission in a 24-month follow-up period. Ethnicity was recorded according to categories defined by the UK Office for National Statistics (11) and documented in participants' health records at the time of first presentation to OASIS or conventional mental health services. The initial diagnosis of first episode psychosis was made by the treating clinician according to ICD-10 criteria (12) and corresponded to the following categories: schizophrenia and related disorders [including patients diagnosed with schizophrenia (F20), delusional disorder (F22) and other schizophrenia-like disorders (F23, F28 and F29)]; schizoaffective disorder (F25); bipolar disorder [including patients receiving a diagnosis of mania (F30) or bipolar disorder

(F31)]; psychotic depression (F32.3, F33.3); drug-related psychosis (F1x.5); other psychoses. A 24-month follow-up period started from the date of referral to OASIS or the conventional service where the diagnosis of first episode psychosis was made. This time period was selected because it permitted assessment of outcomes in the entire sample, and because the outcomes in the first 2 years after illness onset predict the long-term outcomes (13).

Statistical analyses

Statistical analyses were conducted to test the association between predictors and outcomes using STATA 12 software (14) at a significance level of $P < 0.05$. The main exposure was whether the patients with first episode psychosis had been seen by OASIS or by conventional services. The primary outcome was the cumulative duration of hospital admission (in days) during the 24 months of follow-up. This was chosen because in patients with psychosis, the duration of admission indicates the degree of disability and is also the greatest contributor to the costs of clinical care (15). Secondary outcomes were the occurrence of compulsory admission to hospital and the frequency of hospital admission between 2 weeks and 24 months of follow-up. Although it was not possible to investigate differences in the duration of untreated psychosis (DUP), as this variable was not routinely documented in clinical records, an analysis was performed to compare referral-to-diagnosis time between patients seen by OASIS or by conventional services as a proxy measure of DUP.

Age, gender, ethnicity, marital status, employment status, borough of residence, diagnosis and exposure to antipsychotics were included as covariates in multivariable analyses. Where missing data were present in covariates, these were included as explanatory variables in multivariable analyses. Further sensitivity analyses were performed including only participants with complete covariate data to assess the potential impact of missing data.

A sensitivity analysis was performed to test whether there were differences between OASIS data collected before and after 2007. Because the conventional services group included first episode services, supplementary three-way analyses were performed to compare outcomes in patients who presented to the OASIS, first episode services and other SLaM services in order to ascertain any potential differences in outcomes associated with first episode services compared to other SLaM services.

For descriptive analyses, continuous variables were expressed as mean and standard deviation (SD); categorical variables were expressed as frequencies and percentages. Comparison of age distribution between groups was tested using Mann–Witney's *U*-test for two-way analyses and ANOVA for three-way analyses. Chi-square tests were used to compare groups for discrete categorical variables. No differences in *P* values were found when applying Fisher's exact test to contrasts where individual cell frequency was fewer than five. Owing to non-proportionality of hazards, multivariable regression methods were employed at varying periods of follow-up rather than utilising Cox regression for survival analysis. Multiple linear regression models were used to assess the association between engagement by the high-risk service (vs. conventional mental health services) and number of days spent as an in-patient in the first 12 months and 24 months after referral to SLaM, and time to diagnosis from referral to the high-risk service or conventional mental health services. Multivariable binary logistic regression models were used to assess the association of initial management by the high-risk service (vs. conventional mental health services) with compulsory hospital admission under the UK Mental Health Act at 2 weeks, 1 month, 3 months, 6 months, 12 months and 24 months after referral to services. Multivariable Poisson regression models were used to assess the association of initial management by the high-risk service or conventional mental health services with the number of admissions at 2 weeks, 1, 3, 6, 12 and 24 months after referral to services. Poisson regression to analyse number of hospital admissions was employed rather than binary logistic regression for any hospital admission to overcome the ceiling effect encountered by the latter method for individuals with multiple hospital admissions during the follow-up period. Despite a large proportion of zero values for number of hospital admissions, zero-inflated Poisson models were not meaningfully different to standard models (Vuong $P > 0.05$ for all models). For variables with variance greater than mean, negative binomial regression did not yield meaningfully different results to Poisson models (12 months IRR 0.43; 95% CI: 0.32–0.58, 24 months IRR 0.47; 95% CI: 0.36–0.62). For consistency, standard Poisson regression estimates are therefore presented for all variables.

Results

Demographic and diagnostic differences between the samples

The first episode patients that were referred to OASIS were younger and more likely to be male, to belong to an ethnic minority, and to have a schizophrenia spectrum disorder (as opposed to an affective psychosis) than those in the conventional services sample (Table 1).

Primary outcome measure

Multiple linear regression analysis (Table 2) revealed that first episode patients who had been first seen by OASIS spent 17 fewer days in hospital in the 24 months following referral than those first seen by conventional services.

Secondary outcome measures

The median referral-to-diagnosis time for people with first episode psychosis seen by OASIS was shorter than in those presenting to conventional services (Figure 1). Multiple linear regression analysis comparing OASIS with conventional services corroborated this finding [B coefficient −74.5 days, 95% CI: −101.9–(−47.1)]. Multivariable logistic regression analysis (Table 3) showed that among patients presenting with first episode psychosis, those initially seen by OASIS had a reduced likelihood of compulsory hospital admission in the following 24 months (Figure 2). Multivariable Poisson regression analysis also showed that the patients presenting to OASIS had a lower frequency of admission during the follow-up period (Figure 3 and Table 3).

Table 1. Characteristics of patients who were assessed and diagnosed by the high-risk service or conventional mental health services

	High-risk service (n = 164)	Conventional mental health services (n = 2779)	
Mean age (SD)	23.6 (4.88)	25.1 (5.95)	$z = 3.5$ $P < 0.001$
Male gender (%)	112 (68.3%)	1663 (59.8%)	$\chi^2 = 4.6$ $P = 0.03$
Ethnicity (%)			
Black (Black British/Black Caribbean/Black African)	93 (56.7%)	942 (35.6%)	$\chi^2 = 30.0$
Asian	7 (4.3%)	222 (8.4%)	$P < 0.001$
White	51 (31.1%)	1175 (44.5%)	
Other	13 (7.9%)	304 (11.5%)	
Marital status (%)			
Married/cohabiting	12 (7.5%)	275 (11.0%)	$\chi^2 = 2.4$
Divorced/separated	5 (3.1%)	99 (4.0%)	$P = 0.31$
Single	144 (89.4%)	2129 (85.1%)	
Employment status (%)			
Employed	36 (22.9%)	145 (19.1%)	$\chi^2 = 2.4$
Student	31 (19.8%	188 (24.8%)	$P = 0.31$
Unemployed	90 (57.3%)	426 (56.1%)	
Initial diagnosis (%)			
Schizophrenia spectrum	123 (75.0%)	1642 (59.1%)	
Bipolar disorder	8 (4.9%)	142 (5.1%)	
Psychotic depression	6 (3.7%)	312 (11.2%)	$\chi^2 = 21.0$
Schizoaffective disorder	1 (0.6%)	90 (3.2%)	$P = 0.001$
Drug-related psychosis	5 (3.1%)	157 (5.7%)	
Other psychosis	21 (12.8%)	436 (15.7%)	
Borough of residence (%)			
Lambeth	111 (67.7%)	473 (17.0%)	
Southwark	40 (24.4%)	472 (17.0%)	$\chi^2 = 292.9$
Lewisham	11 (6.7%)	442 (15.9%)	$P < 0.001$
Croydon	2 (1.2%)	498 (17.9%)	
Other borough	0 (0.0%)	894 (32.2%)	

SD, standard deviation.

Table 2. Primary outcome: association of prior contact with the high-risk service (n = 164) compared to conventional mental health services (n = 2779) on number of days spent in hospital

	Cumulative change in number of days spent in hospital B coefficient (95% CI)
12 months	−12.7 (−22.5 to (−2.8))
24 months	−17.0 (−33.7 to (−0.3))

Multiple linear regression adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication. Follow-up period commenced from date of referral to the high-risk service or to conventional mental health services.

Sensitivity and supplementary analyses

A sensitivity analysis comparing the data from OASIS collected between 2001 and 2006 and between 2007 and 2011 (Table S1) did not reveal any significant differences. Sensitivity analyses revealed that there were missing covariate data for ethnicity, marital status and employment status, particularly for the conventional service sample (Table S2). However, multivariable analyses including only participants with full covariate data (Table S3a and S3b) did not significantly differ from analyses including missing data as explanatory variables (Table 2 and Table 3). Supplementary three-way analyses excluding first episode services (Table S5a/S6a/S7) revealed similar outcomes to the main analyses. A comparison of first episode services with other conventional services showed an association of first episode services with reduced duration of hospital admission (Table S5b) and compulsory hospital admission, a trend towards reduced number of hospital admissions (Table S6b) and no significant difference in referral-to-diagnosis time (Table S7). A comparison of OASIS with first episode services revealed a non-significant trend towards reduced duration of hospital admission (Table S5c), reduced likelihood of compulsory hospital admission and a significant

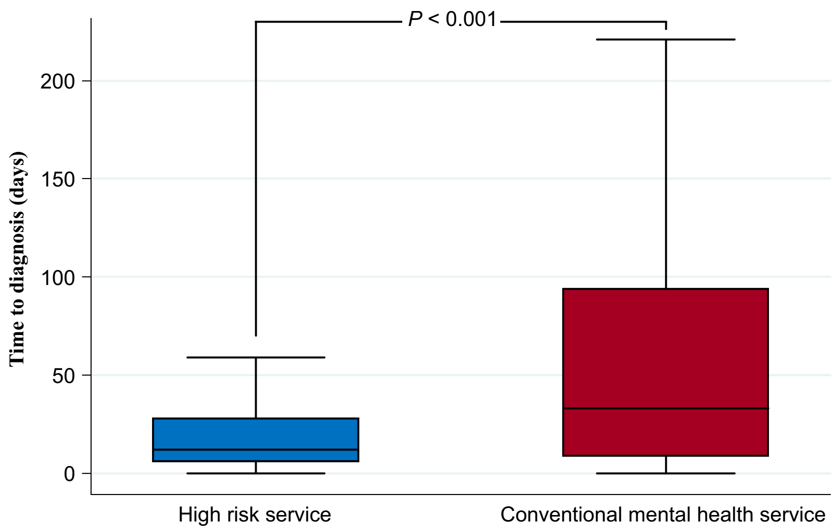


Fig. 1. Time to diagnosis in high-risk service compared to conventional mental health services.

Table 3. Secondary outcomes: association of prior contact with the high-risk service (n = 164) compared to conventional mental health services (n = 2779) on compulsory admission under the UK Mental Health Act and the number of hospital admissions in a given time period

	Any compulsory hospital admission* Odds ratio (95% CI)	Number of hospital admissions† Incidence rate ratio (95% CI)
2 weeks	0.26 (0.10–0.66)	0.13 (0.06–0.30)
1 month	0.29 (0.13–0.62)	0.16 (0.08–0.29)
3 months	0.45 (0.26–0.78)	0.27 (0.18–0.41)
6 months	0.46 (0.28–0.78)	0.34 (0.24–0.48)
12 months	0.53 (0.33–0.84)	0.41 (0.31–0.55)
24 months	0.52 (0.34–0.81)	0.49 (0.39–0.61)

*Multivariable binary logistic regression.
†Multivariable Poisson regression.
All analyses are adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication. Follow-up period commenced from date of referral to the high-risk service or to conventional mental health services.

association with reduced number of hospital admissions (Table S6c) and significant reduction in referral-to-diagnosis time (Table S7).

Discussion

To our knowledge, this is the first study to assess the association of presentation to high-risk services on the outcome of patients referred with a provisional diagnosis of a high-risk state, but found to have first episode psychosis when assessed by the high-risk team. We compared clinical outcomes in this group with those in patients with first episode psychosis who presented to generic mental health services, which included specialised first episode teams. We found that in the 2 years following presentation, patients who initially presented to a high-risk service required fewer hospital admissions were less likely to require compulsory admission and spent fewer days in hospital. These results were independent of differences in age, gender, ethnicity, marital and employment status, borough of residence, psychotic diagnosis and previous exposure to antipsychotic drugs.

The patients who were initially seen by the high-risk team may have been referred to mental healthcare services earlier than they would have been if a high-risk team had not been available. The referrers thought (incorrectly) that these individuals were at high risk for psychosis. However, once they had been fully assessed by a specialist team they were found to be already psychotic. This is relatively common among referrals to high-risk services (1), as it is often difficult to differentiate between the high-risk state and the early stages of first episode psychosis: the symptoms are qualitatively similar, differing only in severity, and the full clinical picture may not emerge until there has been a detailed and lengthy assessment (16). When there is no high-risk service, a patient that is perceived as vulnerable but not frankly psychotic may not be referred to mental health services, as these do not conventionally offer clinical support for this group.

The first episode patients who were referred to OASIS may have been more likely to have been mistaken for being in a high risk as opposed to a psychotic state because their clinical presentation did not conform to that typically encountered in first episode patients. In the UK, there is often a long period between the onset of psychosis and first presentation, by which time the patient is acutely disturbed with severe psychotic symptoms. Patients who present at an earlier stage with less overt psychotic symptoms, or whose symptoms had an insidious rather than an acute onset may be more likely to be misclassified as high risk. Further research on the clinical characteristics of this sub-

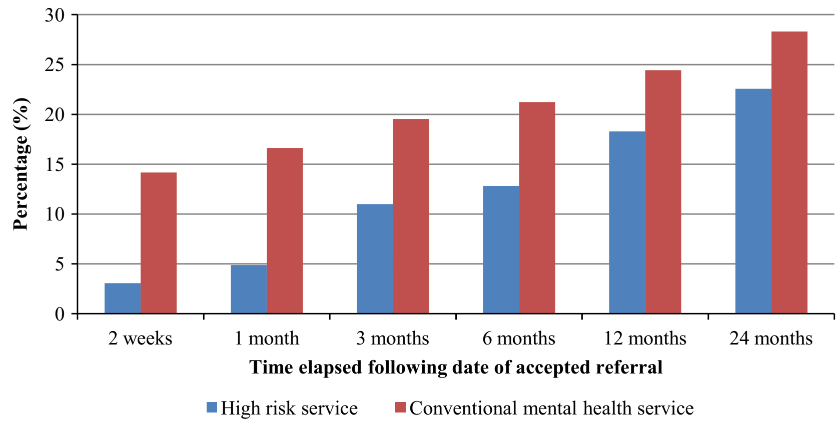


Fig. 2. Cumulative percentage of patients detained under Mental Health Act assessed and diagnosed by the high-risk service (n = 164) compared to conventional mental health services (n = 2779).

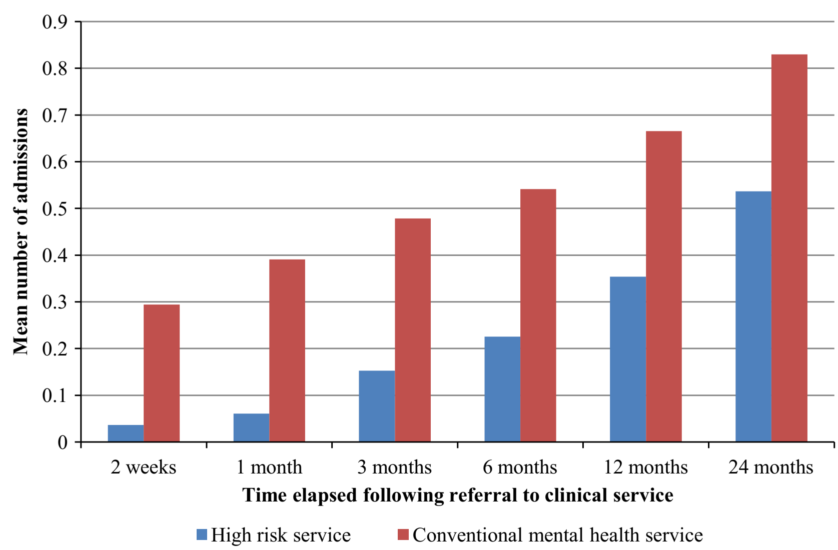


Fig. 3. Mean number of hospital admissions following referral to the high-risk service (n = 164) and conventional mental health services (n = 2779).

group may help to elucidate whether this is the case. Analysis of the demographic features of the two samples in the present study indicated that the patients initially seen by OASIS were significantly younger than those presenting to conventional services. This is consistent with the notion that these patients may have presented at an earlier stage of the first episode.

We also considered whether the better outcomes in the OASIS sample might reflect the presence of sociodemographic features associated with a relatively good prognosis in-patients with first episode psychosis, such as female gender (17), not belonging to an ethnic minority (18) or having a non-schizophrenic psychotic disorder (17). However, comparison of the demographic data from the two samples indicated that the reverse applied: the patients who presented to OASIS were younger, and more likely to be male, from an ethnic minority and to have a schizophreniform psychosis. This may reflect the ethos of high-risk services like OASIS, which mainly operate in a primary care setting, and are designed to be as accessible to patients and referrers as possible. Referrals can be made from any health or non-health agency and via self-referral, and clients are can be seen in their local GP surgery, at home or at a team base in the community. These features may particularly facilitate access to mental health care among patients who are young or who belong to ethnic minority groups (1).

A further potential factor, independent of the nature of clinical features, is that as soon as the patients seen by OASIS had been identified as having first episode psychosis, they were immediately and assertively referred to specialised first episode teams, with an unequivocal diagnosis of first episode psychosis (and not a high-risk state, or any other diagnosis) that was based on a detailed specialist assessment. This ‘fast-track’ form of referral with a clear diagnosis to a closely affiliated first episode team may have resulted in a relatively rapid acceptance of the diagnosis by the receiving team (without the need for further assessment), and relatively quicker initiation of antipsychotic treatment. Delays in accessing specialised services for first episode psychosis can significantly increase the interval between the onset of psychosis and the initiation of antipsychotic treatment, the DUP (19, 20). The greater its duration, the greater the duration of hospitalisation and the risk of rehospitalisation during the first 2 years after referral (21).

Although most of the patients who presented to conventional services were seen by generic mental health teams, about 20% of this sample contacted specialised first episode teams directly. We then tested whether the effect of presenting to a high-risk service was still evident when compared to presenting to a first episode, as opposed to a generic service. Supplementary three-way analyses showed that the there was still a significant reduction in number of hospital admissions, and in the referral-to-diagnosis time in those who presented to the high-risk service, as well as trends towards reduced duration of hospital admission, and reduced rates of compulsory admission. The persistence of differences relative to patients who presented directly to first episode teams suggests that the beneficial effects of presenting to high-risk services are not simply a function of being 'fast-tracked' to a specialised first episode care. Rather, it is consistent with the notion that high-risk services are particularly likely to be referred patients who are in the early stages of the first episode or whose clinical presentation does not immediately suggest that they are psychotic.

In the present study, patients who were initially seen by a high-risk service had fewer hospital admissions and spent 17 fewer days in hospital within the first 2 years than patients who presented to conventional services. Hospital admissions are the single largest contributor to the direct costs associated with the care of schizophrenia (15). In the UK, the average cost of a night in a psychiatric bed is £350 GBP (15), and an average cost of £12 198 GBP per admission has been estimated (22). We also found that the patients who were initially seen by a high-risk service were less likely to require a compulsory admission under the Mental Health Act. Compulsory admissions are usually longer than voluntary admissions and are associated with higher direct costs (22). In addition, the experience of compulsory admission can be a negative one for both the patient and their family: this may have an adverse effect on the patient's subsequent engagement with mental health services and their adherence to treatment, and is associated with an increased risk of further compulsory admissions (23).

This was an observational study, and patients were not randomly assigned to the two samples that were compared. However, as high-risk services are not designed to manage first episode patients, a study in which patients were randomly allocated to high risk and conventional teams would be impractical and ethically problematic. In our study, we investigated variations in clinical outcomes in a single provider of mental health care (SLaM). Another approach which could be investigated in future studies is to compare outcomes in different providers of mental health care depending on whether or not they provide high-risk clinical services. However, such an approach would not necessarily overcome these limitations because of heterogeneity due to differences in characteristics between different mental healthcare providers. We performed a retrospective assessment of the data and used information that had been entered by clinicians in the patients' records. Data completion was satisfactory for all the relevant outcomes, and sensitivity analysis did not show significant differences with respect to missing data. However, the information in the clinical records did not include a standardised measure of illness severity at the time of the first episode psychosis diagnosis, and we were therefore unable to control for this potentially confounding factor in the analysis. However, we chose not to employ a propensity score approach as there is evidence that this method does not overcome the limitation of residual confounding (24). We were also unable to control for all treatments received from the first episode psychosis diagnosis over the follow-up time. However, the use of antipsychotic exposure as proxy index of treatment offered after the first episode psychosis diagnosis did not affect our findings.

This study provides the first evidence that services designed for people at high risk of psychosis may be associated with better outcomes in patients who are already psychotic, but were referred because they were thought to be at high risk. This may result from the referral of patients at a relatively early stage of the first episode and from the fast-tracking of these patients to specialised first episode services. Both are likely to reduce the interval between the onset of psychosis and the initiation of antipsychotic treatment.

Acknowledgement

We thank all the OASIS and SLaM patients.

Declaration of interest

The CRIS team (H.S., M.B., R.S.) have received research funding from Roche; Pfizer; Johnson & Johnson; and Lundbeck. P.M. has received research funding from Janssen; Sunovion; GW Pharmaceuticals; and Roche. Funding organisations had no role in the collection, management, analysis and interpretation of the data; and the preparation, review or approval of the manuscript.

Ethical approval

Ethical approval for the study was obtained from the Institutional Review Board of the SLaM Psychosis Clinical Academic Group (CAG) for collection and analysis of data on patients presenting to the high-risk service (OASIS) and Oxfordshire REC C (Ref: 08/H0606/71 + 5) for collection and analysis of data from the BRC Case Register for patients presenting to conventional mental health services.

Funding

This work was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London; Instituto de Salud Carlos III, Spanish Ministry of Economy and Competitiveness and Fundación Alicia Koplowitz (to C.M.D-C.); UK Medical Research Council Clinical Research Training Fellowship (MR/K002813/1 to R.P.).

References

1. Fusar-Poli P, Byrne M, Badger S, Valmaggia LR, McGuire PK. Outreach and support in South London (OASIS), 2001–2011: ten years of early diagnosis and treatment for young individuals at high clinical risk for psychosis. Eur Psychiatry 2013;**28**:315–326.
2. Fusar-Poli P, Borgwardt S, Bechdolf A et al. The psychosis high-risk state: a comprehensive state-of-the-art review. JAMA Psychiatry 2013;**70**:107–120.
3. Stafford MR, Jackson H, Mayo-Wilson E, Morrison AP, Kendall T. Early interventions to prevent psychosis: systematic review and meta-analysis. BMJ 2013;**346**:f185.
4. Valmaggia LR, McCrone P, Knapp M et al. Economic impact of early intervention in people at high risk of psychosis. Psychol Med 2009;**39**:1617–1626.
5. Nelson B, Yung AR. When things are not as they seem: detecting first-episode psychosis upon referral to ultra high risk ('prodromal') clinics. Early Interv Psychiatry 2007;**1**:208–211.
6. Perkins DO, Gu H, Boteva K, Lieberman JA. Relationship between duration of untreated psychosis and outcome in first-episode schizophrenia: a critical review and meta-analysis. Am J Psychiatry 2005;**162**:1785–1804.
7. Stewart R, Soremekun M, Perera G et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry 2009;**9**:51.
8. Kempton MJ, Bonoldi I, Valmaggia L, McGuire P, Fusar-Poli P. Speed of Psychosis Progression in People at Ultra-High Clinical Risk: A Complementary Meta-analysis. JAMA Psychiatry 2015;**72**:622–623.
9. Valmaggia LR, Byrne M, Day F et al. Duration of untreated psychosis and need for admission in patients who engage with mental health services in the prodromal phase. Br J Psychiatry 2015;**207**:130–134.
10. Mental Health Act. Great Britain: London: The Stationery Office, 2007. http://www.legislation.gov.uk/ukpga/2007/12/contents
11. Office for National Statistics. Ethnic Group. London. http://www.ons.gov.uk/ons/guide-method/measuring-equality/equality/ethnic-nat-identity-religion/ethnic-group/index.html
12. *International statistical classification of diseases and related health problems*. 10th revis. World Health Organization, 2004
13. Harrison G, Hopper K, Craig T et al. Recovery from psychotic illness: a 15-and 25-year international follow-up study. Br J Psychiatry 2001;**178**:506–517.
14. StataCorp. Stata statistical software: release 12. College Station, TX: StataCorp LP, 2011.
15. Knapp M, Andrew A, McDaid D et al. Investing in recovery: making the business case for effective interventions for people with schizophrenia and psychosis. London Rethink Ment Illn 2014; http://eprints.lse.ac.uk/56773/.
16. Zimbrón J, Ruiz de Azúa S, Khandaker GM et al. Clinical and sociodemographic comparison of people at high-risk for psychosis and with first-episode psychosis. Acta Psychiatr Scand 2013;**127**:210–216.
17. Clemmensen L, Vernal DL, Steinhausen H-C. A systematic review of the long-term outcome of early onset schizophrenia. BMC Psychiatry 2012;**12**:150.
18. Coid JW, Kirkbride JB, Barker D et al. Raised incidence rates of all psychoses among migrant groups: findings from the East London first episode psychosis study. Arch Gen Psychiatry 2008;**65**:1250–1258.
19. Birchwood M, Connor C, Lester H et al. Reducing duration of untreated psychosis: care pathways to early intervention in psychosis services. Br J Psychiatry 2013;**203**:58–64.
20. Morgan C, Abdul-Al R, Lappin JM et al. Clinical and social determinants of duration of untreated psychosis in the AESOP first-episode psychosis study. Br J Psychiatry 2006;**189**:446–452.
21. Penttilä M, Miettunen J, Koponen H et al. Association between the duration of untreated psychosis and short- and long-term outcome in schizophrenia within the Northern Finland 1966 Birth Cohort. Schizophr Res 2013;**143**:3–10.
22. Andrews A, Knapp M, McCrone P, Parsonage M, Trachtenberg M. Effective interventions in schizophrenia the economic case: a report prepared for the Schizophrenia Commission. London Rethink Ment Illn 2012. http://eprints.lse.ac.uk/47406/.
23. Jaeger S, Pfiffner C, Weiser P et al. Long-term effects of involuntary hospitalization on medication adherence, treatment engagement and perception of coercion. Soc Psychiatry Psychiatr Epidemiol 2013;**48**:1787–1796.
24. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. BMJ 2013;**347**:f6409.

Supporting Information

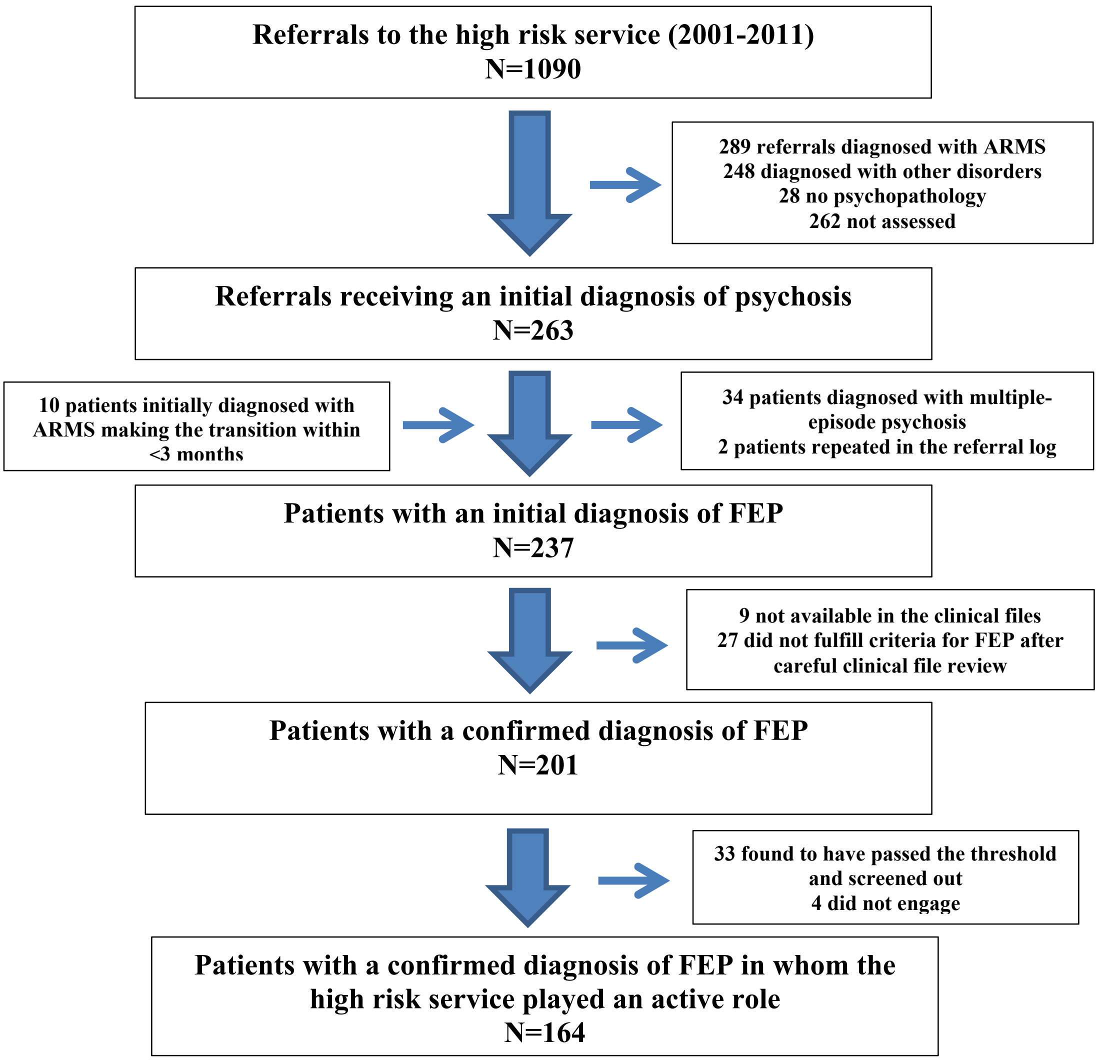
Additional Supporting Information may be found in the online version of this article:

Figure S1. Sample selection for the high risk service.

Table S1. Characteristics of patients referred to the high risk service in 2001–2006 compared to patients referred in 2007–2011.
Table S2. Characteristics of patients who were assessed and diagnosed by the high risk service or conventional mental health services including missing covariate data.
Table S3 (a). Primary outcome: association of prior contact with the high risk service ($n = 164$) compared to conventional mental health services ($n = 2779$) on number of days spent in hospital. Analysis including only participants with full covariate data. (b). Secondary outcomes: association of prior contact with the high risk service ($n = 164$) compared to conventional mental health services ($n = 2779$) on compulsory admission under the UK Mental Health Act and the number of hospital admissions in a given time period. Analysis including only participants with full covariate data.
Table S4. Characteristics of patients who were assessed and diagnosed by the high risk service, first episode service or to other conventional mental health services.
Table S5 (a) Association of prior contact with the high risk service ($n = 164$) compared to other conventional mental health services, not including first episode services ($n = 2284$) on number of days spent in hospital. (b) Association of prior contact with the first episode service ($n = 495$) compared to other conventional mental health services ($n = 2284$) on number of days spent in hospital. (c) Association of prior contact with the high risk service ($n = 164$) compared to the first episode service ($n = 495$) on number of days spent in hospital.
Table S6 (a) Association of prior contact with the high risk service ($n = 164$) compared to other conventional mental health services, not including first episode services ($n = 2284$) on compulsory admission under the UK Mental Health Act and the number of hospital admissions in a given time period. (b) Association of prior contact with the first episode service ($n = 495$) compared to other conventional mental health services ($n = 2284$) on compulsory admission under the UK Mental Health Act and the number of hospital admissions in a given time period. (c) Association of prior contact with the high risk service ($n = 164$) compared to the first episode service ($n = 495$) on compulsory admission under the UK Mental Health Act and the number of hospital admissions in a given time period.
Table S7. Association of prior contact with the high risk service ($n = 164$), first episode service ($n = 495$) and other conventional mental health services ($n = 2284$) on referral-to-diagnosis time from referral to services.

SERVICES FOR PEOPLE AT HIGH RISK IMPROVE OUTCOME IN PATIENTS WITH FIRST EPISODE PSYCHOSIS – SUPPLEMENTARY MATERIAL

eFigure 1: SAMPLE SELECTION FOR THE HIGH RISK SERVICE



Abbreviations: ARMS= At Risk Mental State, FEP= First Episode Psychosis.

eTable1: Characteristics of patients referred to the high risk service in 2001-2006 compared to patients referred in 2007-2011

	High risk service 2001-2006 (n=72)	High risk service 2007-2011 (n=92)	
Mean age (SD)	23.7 (5.02)	23.4 (4.80)	z=0.13 p=0.90
Male gender (%)	53 (73.6%)	59 (64.1%)	χ^2=1.68 p=0.20
Ethnicity (%)			
Black (Black British/ Black Caribbean/ Black African)	44 (61.1%)	49 (53.3%)	
Asian	5 (6.9%)	2 (2.2%)	χ^2=7.04
White	21 (29.2%)	30 (32.6%)	p=0.07
Other	2 (2.8%)	11 (12.0%)	
Marital status (%)			
Married/cohabiting	3 (4.2%)	9 (10.1%)	
Divorced/separated	3 (4.2%)	2 (2.3%)	χ^2=2.43 p=0.30
Single	66 (91.7%)	78 (87.6%)	
Employment status (%)			
Employed	13 (18.6%)	23 (26.4%)	χ^2=1.70
Student	16 (22.9%)	15 (17.2%)	p=0.43
Unemployed	41 (58.6%)	49 (56.3%)	
Initial diagnosis (%)			
- Schizophrenia-like	57 (79.2%)	66 (71.7%)	
- Bipolar disorder	3 (4.2%)	5 (5.4%)	
- Psychotic depression	2 (2.8%)	4 (4.4%)	χ^2=2.96
- Schizoaffective disorder	1 (1.4%)	0 (0%)	p=0.71
- Drug-related psychosis	2 (2.8%)	3 (3.3%)	
- Other psychosis	7 (9.7%)	14 (15.2%)	

eTable 2: Characteristics of patients who were assessed and diagnosed by the high risk service or conventional mental health services including missing covariate data

	High risk service (n=164)	Conventional mental health services (n=2779)	
Mean age (SD)	23.6 (4.88)	25.1 (5.95)	z=3.5 $p<0.001$
Male gender (%)	112 (68.3%)	1663 (59.8%)	$\chi^2=4.6$ p=0.03
Ethnicity (%)			
Black (Black British/ Black Caribbean/ Black African)	93 (56.7%)	942 (33.9%)	
Asian	7 (4.3%)	222 (8.0%)	$\chi^2=39.6$
White	51 (31.1%)	1175 (45.3%)	$p<0.001$
Other	13 (7.9%)	304 (10.9%)	
Not recorded	0 (0.0%)	136 (4.9%)	
Marital status (%)			
Married/cohabiting	12 (7.3%)	275 (9.9%)	
Divorced/separated	5 (3.1%)	99 (3.6%)	$\chi^2=14.4$
Single	144 (87.8%)	2129 (76.6%)	p=0.002
Not recorded	3 (1.8%)	276 (9.9%)	
Employment status (%)			
Employed	36 (22.0%)	145 (5.2%)	$\chi^2=344.6$
Student	31 (18.9%	188 (6.8%)	$p<0.001$
Unemployed	90 (54.9%)	426 (15.3%)	
Not recorded	7 (4.3%)	2020 (72.7%)	
Initial diagnosis (%)			
Schizophrenia-spectrum	123 (75.0%)	1642 (59.1%)	
Bipolar disorder	8 (4.9%)	142 (5.1%)	
Psychotic depression	6 (3.7%)	312 (11.2%)	$\chi^2=21.0$
Schizoaffective disorder	1 (0.6%)	90 (3.2%)	p=0.001
Drug-related psychosis	5 (3.1%)	157 (5.7%)	
Other psychosis	21 (12.8%)	436 (15.7%)	
Borough of residence (%)			
Lambeth	111 (67.7%)	473 (17.0%)	
Southwark	40 (24.4%)	472 (17.0%)	$\chi^2=292.9$
Lewisham	11 (6.7%)	442 (15.9%)	$p<0.001$
Croydon	2 (1.2%)	498 (17.9%)	
Other borough	0 (0.0%)	894 (32.2%)	

eTable 3a: Primary outcome: association of prior contact with the high risk service (n=164) compared to conventional mental health services (n=2779) on number of days spent in hospital. Analysis including only participants with full covariate data.

	Cumulative change in number of days spent in hospital B coefficient (95% CI)
12 months	-15.6 days (95% CI -25.2 to -6.0)
24 months	-22.2 days (95% CI -38.5 to -6.0)

Multiple linear regression adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication
Follow-up period commenced from date of referral to the high risk service or to conventional mental health services

eTable 3b: Secondary outcomes: association of prior contact with the high risk service (n=164) compared to conventional mental health services (n=2779) on compulsory admission under the UK Mental Health Act and the number of hospital admissions in a given time period. Analysis including only participants with full covariate data.

	Any compulsory hospital admission* Odds ratio (95% CI)	Number of hospital admissions** Incidence rate ratio (95% CI)
2 weeks	0.16 (0.05 to 0.48)	0.12 (0.05 to 0.29)
1 month	0.18 (0.08 to 0.45)	0.15 (0.08 to 0.30)
3 months	0.32 (0.17 to 0.61)	0.29 (0.19 to 0.44)
6 months	0.33 (0.18 to 0.60)	0.36 (0.25 to 0.52)
12 months	0.40 (0.24 to 0.69)	0.43 (0.32 to 0.58)
24 months	0.42 (0.26 to 0.70)	0.48 (0.38 to 0.62)

**Multivariable binary logistic regression*
***Multivariable Poisson regression*
All analyses are adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication
Follow-up period commenced from date of referral to the high risk service or to conventional mental health services

eTable 4: Characteristics of patients who were assessed and diagnosed by the high risk service, first episode service or to other conventional mental health services

	High risk service (n=164)	First episode service (n=495)	Other conventional mental health services (n=2284)	
Mean age (SD)	23.5 (4.88)	23.8 (5.16)	25.35 (6.07)	ANOVA F=19.6 P<0.001
Male gender (%)	112 (68.3%)	338 (68.3%)	1325 (58.0%)	χ^2=22.5 p<0.001
Ethnicity (%)				
Black (Black British/ Black Caribbean/ Black African)	93 (56.7%)	234 (49.0%)	708 (32.7%)	χ^2=90.3 p<0.001
Asian	7 (4.3%)	36 (7.5%)	186 (8.6%)	
White	51 (31.1%)	142 (29.7%)	1033 (47.7%)	
Other	13 (7.9%)	66 (13.8%)	238 (11.0%)	
Marital status (%)				
Married/cohabiting	12 (7.5%)	37 (7.7%)	238 (11.8%)	χ^2=9.5 p=0.05
Divorced/separated	5 (3.1%)	22 (4.6%)	77 (3.8%)	
Single	144 (89.4%)	423 (87.8%	1706 (84.4%)	
Employment status (%)				
Employed	36 (22.9%)	35 (21.1%)	110 (18.6%)	χ^2=5.2 p=0.26
Student	31 (19.8%	33 (19.9%)	155 (26.1%)	
Unemployed	90 (57.3%)	98 (59.0%)	328 (55.3%)	
Initial diagnosis (%)				
Schizophrenia-spectrum	123 (75.0%)	319 (64.4%)	1323 (57.9%)	χ^2=61.1 p<0.001
Bipolar disorder	8 (4.9%)	17 (3.4%)	125 (5.5%)	
Psychotic depression	6 (3.7%)	28 (5.7%)	284 (12.4%)	
Schizoaffective disorder	1 (0.6%)	8 (1.6%)	822 (3.6%)	
Drug-related psychosis	5 (3.1%)	21 (4.2%)	136 (6.0%)	
Other psychosis	21 (12.8%)	102 (20.6%)	334 (14.6%)	
Borough of residence (%)				
Lambeth	111 (67.7%)	171 (34.6%)	302 (13.2%)	χ^2=461.7 p<0.001
Southwark	40 (24.4%)	107 (21.6%)	365 (16.0%)	
Lewisham	11 (6.7%)	73 (14.8%)	369 (16.2%)	
Croydon	2 (1.2%)	75 (15.2%)	423 (18.5%)	
Other borough	0 (0.0%)	69 (13.9%)	825 (36.1%)	

eTable 5a: Association of prior contact with the high risk service (n=164) compared to other conventional mental health services, not including first episode services (n=2284) on number of days spent in hospital.

	High risk service Cumulative change in number of days spent in hospital B coefficient (95% CI)
12 months	-15.3 (-13.3 to -1.6)
24 months	-20.7 (-37.8 to -3.7)

Multiple linear regression adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication
Follow-up period commenced from date of referral to the high risk service or to other conventional mental health services

eTable 5b: Association of prior contact with the first episode service (n=495) compared to other conventional mental health services (n=2284) on number of days spent in hospital.

	First episode service Cumulative change in number of days spent in hospital B coefficient (95% CI)
12 months	-7.4 (-13.3 to -1.6)
24 months	-10.5 (-20.5 to -0.6)

Multiple linear regression adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication
Follow-up period commenced from date of referral to the first episode service or to other conventional mental health services

eTable 5c: Association of prior contact with the high risk service (n=164) compared to the first episode service (n=495) on number of days spent in hospital.

	High risk service Cumulative change in number of days spent in hospital B coefficient (95% CI)
12 months	-7.9 (-18.4 to 2.7)
24 months	-10.2 (-28.1 to 7.7)

Multiple linear regression adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication
Follow-up period commenced from date of referral to the high risk service or to the first episode service

eTable 6a: Association of prior contact with the high risk service (n=164) compared to other conventional mental health services, not including first episode services (n=2284) on compulsory admission under the UK Mental Health Act and the number of hospital admissions in a given time period.

	Any compulsory hospital admission* Odds ratio (95% CI)	Number of hospital admissions** Incidence rate ratio (95% CI)
2 weeks	0.20 (0.08 to 0.53)	0.13 (0.06 to 0.28)
1 month	0.23 (0.10 to 0.50)	0.15 (0.08 to 0.28)
3 months	0.37 (0.21 to 0.65)	0.26 (0.17 to 0.39)
6 months	0.39 (0.23 to 0.66)	0.33 (0.23 to 0.46)
12 months	0.46 (0.28 to 0.73)	0.40 (0.30 to 0.53)
24 months	0.48 (0.31 to 0.75)	0.47 (0.37 to 0.59)

**Multivariable binary logistic regression*
***Multivariable Poisson regression*
All analyses are adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication
Follow-up period commenced from date of referral to the high risk service or to other conventional mental health services

eTable 6b: Association of prior contact with the first episode service (n=495) compared to other conventional mental health services (n=2284) on compulsory admission under the UK Mental Health Act and the number of hospital admissions in a given time period.

	Any compulsory hospital admission* Odds ratio (95% CI)	Number of hospital admissions** Incidence rate ratio (95% CI)
2 weeks	0.46 (0.33 to 0.65)	0.85 (0.70 to 1.02)
1 month	0.52 (0.39 to 0.70)	0.84 (0.72 to 0.99)
3 months	0.61 (0.46 to 0.80)	0.89 (0.77 to 1.02)
6 months	0.62 (0.48 to 0.80)	0.88 (0.77 to 1.01)
12 months	0.69 (0.54 to 0.88)	0.91 (0.80 to 1.03)
24 months	0.81 (0.64 to 1.02)	0.91 (0.82 to 1.02)

**Multivariable binary logistic regression*
***Multivariable Poisson regression*
All analyses are adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication
Follow-up period commenced from date of referral to the first episode service or to other conventional mental health services

eTable 6c: Association of prior contact with the high risk service (n=164) compared to the first episode service (n=495) on compulsory admission under the UK Mental Health Act and the number of hospital admissions in a given time period.

	Any compulsory hospital admission* Odds ratio (95% CI)	Number of hospital admissions** Incidence rate ratio (95% CI)
2 weeks	0.44 (0.16 to 1.17)	0.15 (0.06 to 0.34)
1 month	0.44 (0.20 to 0.98)	0.17 (0.09 to 0.33)
3 months	0.61 (0.34 to 1.09)	0.29 (0.19 to 0.45)
6 months	0.62 (0.36 to 1.08)	0.37 (0.26 to 0.53)
12 months	0.66 (0.40 to 1.08)	0.44 (0.33 to 0.59)
24 months	0.59 (0.38 to 0.94)	0.51 (0.40 to 0.65)

**Multivariable binary logistic regression*
***Multivariable Poisson regression*
All analyses are adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication
Follow-up period commenced from date of referral to the high risk service or to the first episode service

eTable 7: Association of prior contact with the high risk service (n=164), first episode service (n=495) and other conventional mental health services (n=2284) on referral-to-diagnosis time from referral to services.

	Change in referral-to-diagnosis time B coefficient (95% CI)
High risk vs. other conventional mental health services	-70.3 days (-98.3 to -42.3)
First episode vs. other conventional mental health services	12.0 days (-4.3 to 28.4)
High risk vs. first episode services	-82.3 days (-111.7 to -52.9)

Multiple linear regression adjusted for age, gender, ethnicity, marital status, employment status, diagnosis, borough of residence and whether receiving antipsychotic medication

3.2 Supplementary methods

3.21 SQL data extraction

Data on patients with FEP receiving care from the high risk service were obtained by manual extraction from ePJS software. Data on patients with FEP receiving care from conventional mental health services was obtained by data extraction from the BRC Case Register using a series of SQL scripts described below.

```
USE [SQLCRIS_User]
select brcid,accepted_date,spell
from (
        select distinct brcid,accepted_date,ROW_NUMBER()
        over (partition by brcid order by accepted_date) spell
        from [brhnsql094].[SQLCRIS].[dbo].[Referral] r
        where
        Accepted_Date>='1 Jan 1900'
        and Referral_Status_ID not like 'rejected'
    ) a
where  Accepted_Date between '1 jan 2007' and '31 dec 2012' and spell=1
```

This script selected patients presenting to SLaM for the first time between 1st January 2007 and 31st December 2012. The method involves a sub-query within the primary SELECT statement. The sub-query selects data from the Referral table whose currency is individual referrals. Therefore, each patient (represented by the BRCID) may have many referrals if they have been referred to SLaM multiple times. The command, “over (partition by brcid order by accepted_date) spell” creates an additional column, “spell”, which consists of integers which count upwards for each successive referral per patient. The “Accepted_Date>='1 Jan 1900'” clause is a data quality statement to ensure that rows with null values (i.e. missing data) are not included in the data extraction. The “Referral_Status_ID not like 'rejected'” clause adds a filter to exclude any referrals which were rejected. Finally, the “where Accepted_Date between '1 jan 2007' and '31 dec 2012' and spell=1” clause outside of the sub-query selects patients who were referred to SLaM for the first time within the specified date range.

The output of this query was saved as a view in the following location: “[SQLCRIS_User].[RPatel].[rp_firstreferral_2007_2012]” from which the next script was applied.

```
USE [SQLCRIS_User]
SELECT     a.brcid, accepted_date, b.diagnosis_date
Dateoffirstpsychosisdiagnosis, b.primary_diagnosis Firstpsychosisdiagnosis
FROM         RPatel.rp_firstreferral_2007_2012 a LEFT JOIN
  (SELECT     brcid, diagnosis_date, primary_diagnosis
  FROM          (SELECT     brcid, primary_diagnosis, diagnosis_date,
source_table, ROW_NUMBER() OVER (partition BY brcid
  ORDER BY diagnosis_date) AS dn
  FROM          SQLCrisImport.dbo.diagnosis_combined
  WHERE       (primary_diagnosis LIKE '%psychotic%' OR
  primary_diagnosis LIKE '%psychosis%' OR
  primary_diagnosis LIKE '%with psyc%' OR
  primary_diagnosis LIKE '%schizophreni%' OR
  primary_diagnosis LIKE '%scizophreni%' OR
  primary_diagnosis LIKE '%schizotyp%' OR
  primary_diagnosis LIKE '%scizotyp%' OR
  primary_diagnosis LIKE '%delusion%' OR
  primary_diagnosis LIKE '%hallucin%' OR
  primary_diagnosis LIKE '%thought disorder%' OR
  primary_diagnosis LIKE '%first rank%' OR
  primary_diagnosis LIKE '%schizoaffect%' OR
  primary_diagnosis LIKE '%scizoaffect%' OR
  primary_diagnosis LIKE '%mania%' OR
  primary_diagnosis LIKE '%manic%' OR
  primary_diagnosis LIKE '%f20%' OR
  primary_diagnosis LIKE '%f21%' OR
  primary_diagnosis LIKE '%f22%' OR
  primary_diagnosis LIKE '%f23%' OR
  primary_diagnosis LIKE '%f24%' OR
  primary_diagnosis LIKE '%f25%' OR
  primary_diagnosis LIKE '%f28%' OR
  primary_diagnosis LIKE '%f29%' OR
  primary_diagnosis LIKE '%f30%' OR
  primary_diagnosis LIKE '%f31.2%' OR
  primary_diagnosis LIKE '%f32.3%' OR
  primary_diagnosis LIKE '%f33.3%') AND primary_diagnosis NOT LIKE
'%without psyc%' AND
```

```
  primary_diagnosis NOT LIKE '%personality%' AND primary_diagnosis NOT LIKE
'%trichotil%' AND
  primary_diagnosis NOT LIKE '%kleptomani%' AND primary_diagnosis NOT LIKE
'%hypomani%' AND diagnosis_date BETWEEN
  '01-jan-2007' AND '31-dec-2012') AS d
WHERE dn = 1) b ON a.BrcId = b.brcid
WHERE b.diagnosis_date IS NOT NULL AND (b.diagnosis_date > a.Accepted_Date)
```

This script selects all patients previously defined as being referred to SLaM for the first time between 1st January 2007 and 31st December 2012 who have a recorded diagnosis of a psychotic disorder after the date of their SLaM referral being accepted. The criteria for defining a psychotic disorder are defined in the where statement and includes text and ICD-10 code filters for schizophrenia spectrum disorders, mania, psychotic depression or other psychotic disorder.

The output of this query was saved as a view in the following location: "[SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_2007_2012]" from which the next script was applied.

```
USE [SQLCRIS_User]
select
        a.[brcid],
    a.[accepted_date],
    h.Accepted_Date Dateofclosestteamepisode,
    [Dateoffirstpsychosisdiagnosis],
    [Firstpsychosisdiagnosis] Firstpsychosisdiagnosisraw,
        case
                when (
                        Firstpsychosisdiagnosis like '%schizophreni%' or
                        Firstpsychosisdiagnosis like '%scizophreni%' or
                        Firstpsychosisdiagnosis like '%schizotyp%' or
                        Firstpsychosisdiagnosis like '%scizotyp%' or
                        Firstpsychosisdiagnosis like '%f20%' or
                        Firstpsychosisdiagnosis like '%f21%' or
                        Firstpsychosisdiagnosis like '%f22%' or
                        Firstpsychosisdiagnosis like '%f23%' or
                        Firstpsychosisdiagnosis like '%f24%' or
                        Firstpsychosisdiagnosis like '%f28%' or
                        Firstpsychosisdiagnosis like '%f29%')
                then 'F2xSchizophrenia'
```

```
                when (
                        Firstpsychosisdiagnosis like '%schizoaffect%' or
                        Firstpsychosisdiagnosis like '%scizoaffect%' or
                        Firstpsychosisdiagnosis like '%f25%')
                then 'Schizoaffective'

                when (
                        Firstpsychosisdiagnosis like '%f30%' or
                        Firstpsychosisdiagnosis like '%f31%' or
                        Firstpsychosisdiagnosis like '%manic%' or
                        Firstpsychosisdiagnosis like '%mania%' or
                        Firstpsychosisdiagnosis like '%bipolar%' or
                        Firstpsychosisdiagnosis like '%bpad%' or
                        Firstpsychosisdiagnosis like '%affective disorder%'
or
                        Firstpsychosisdiagnosis like '%mixed affective%')
and
                        Firstpsychosisdiagnosis not like
'%trichotillomania%' and
                        Firstpsychosisdiagnosis not like '%kleptomania%'
                then 'Bipolar'

                when (
                        Firstpsychosisdiagnosis like '%psychosis%' or
                        Firstpsychosisdiagnosis like '%psychotic%' or
                        Firstpsychosisdiagnosis like '%with psyc%' or
                        Firstpsychosisdiagnosis like '%f32.3%' or
                        Firstpsychosisdiagnosis like '%f33.3%') and
                        (Firstpsychosisdiagnosis like '%depress%' or
                        Firstpsychosisdiagnosis like '%f32%' or
                        Firstpsychosisdiagnosis like '%f33%') and
                        Firstpsychosisdiagnosis not like '%without%'
                then 'PsychoticDepression'

                when (
                        Firstpsychosisdiagnosis like '%drug%' or
                        Firstpsychosisdiagnosis like '%alcohol%' or
                        Firstpsychosisdiagnosis like '%opioid%' or
                        Firstpsychosisdiagnosis like '%opiate%' or
                        Firstpsychosisdiagnosis like '%cannabi%' or
                        Firstpsychosisdiagnosis like '%benzo%' or
                        Firstpsychosisdiagnosis like '%hallucinogen%' or
                        Firstpsychosisdiagnosis like '%cocaine%' or
                        Firstpsychosisdiagnosis like '%cannabis%' or
                        Firstpsychosisdiagnosis like '%f10%' or
                        Firstpsychosisdiagnosis like '%f11%' or
                        Firstpsychosisdiagnosis like '%f12%' or
                        Firstpsychosisdiagnosis like '%f13%' or
```

```
                       Firstpsychosisdiagnosis like '%f14%' or
                       Firstpsychosisdiagnosis like '%f15%' or
                       Firstpsychosisdiagnosis like '%f16%' or
                       Firstpsychosisdiagnosis like '%f17%' or
                       Firstpsychosisdiagnosis like '%f18%' or
                       Firstpsychosisdiagnosis like '%f19%')
               then 'F1x.5DrugPsyc'

               else 'OtherPsychosis'

       end as Firstpsychosisdiagnosisrecode

  FROM [SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_2007_2012] a

left join SQLCris.dbo.Team_episode h on h.CN_Doc_ID=
(select top 1 CN_Doc_ID from SQLCris.dbo.Team_episode h1 where
h1.BrcId=a.BrcId and (h1.Accepted_Date <=a.Dateoffirstpsychosisdiagnosis)
order by h1.Accepted_Date desc)
```

This script recodes the psychotic disorder diagnosis into separate categories in the CASE WHEN statement and also includes a table join to identify the date that a patient was first accepted by the SLaM team which recorded the psychotic disorder diagnosis. This date may occur after the date of the initial referral to SLaM being accepted as patients may have had previous episodes of care relating to non-psychotic disorders prior to being diagnosed with psychosis. The method for identifying the date of first acceptance by a SLaM team (coded as “Dateofclosestteamepisode”), involves a left table join with a subquery. In this example, table “a” (“[SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_2007_2012]”) is joined to a subset of table “h” (“SQLCris.dbo.Team_episode”) in which the first episode of care is chosen (represented by the “top 1 CN_Doc_ID”) in which before the date of being diagnosed with a psychotic disorder occurs (“h1.Accepted_Date <=a.Dateoffirstpsychosisdiagnosis”). This technique allows for data from tables of different currencies which have a many-to-one relationship (in this case, table “a” is at patient level whereas table “h” is at team episode level and patients may have multiple team episodes) to be joined through a one-to-one relationship.

The output of this query was saved as a view in the following location:

"[SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_recode_2007_2012]" from which the next script was applied.

```
USE [SQLCRIS_User]
select
        r.[brcid],
        [accepted_date] Dateoffirstreferral,
        [Dateofclosestteamepisode],
        [Dateoffirstpsychosisdiagnosis],
        [Firstpsychosisdiagnosisraw],
        [Firstpsychosisdiagnosisrecode],
    la_name Borough,
        DATEDIFF(day,[accepted_date],[Dateoffirstpsychosisdiagnosis])
DiagnosticDelayInDays,
        floor((datediff (day, p.cleaneddateofbirth , cast (accepted_date as
datetime))/365)) age,
        p.Gender_ID  Gender,
        p.ethnicitycleaned ethnicity,
        p.Marital_Status_ID marital_status,
        p.Employment_ID Employment_status,
        p.Housing_Status Accommodation_status,
        case when (
                select distinct brcid
                from
SQLCRIS_User.RPatel.rp_firstreferral_psychosis_teamwardbefore_2007_2012
                where (
                        Location_Name like '%COAST%' or
                        Location_Name like '%Leo Community%' or
                        Location_Name like '%Leo Crisis%' or
                        Location_Name like '%Lewisham Early%' or
                        Location_Name like '%STEP%')
                        and
                        brcid=r.brcid
                ) IS not null then 1 else 0 end PriorEISCommunity,
        case when (
                select distinct brcid
                from
SQLCRIS_User.RPatel.rp_firstreferral_psychosis_teamwardbefore_2007_2012
                where (
                        Location_Name like '%Leo Unit%')
                        and
                        brcid=r.brcid
                ) IS not null then 1 else 0 end PriorEISInpatientLEO,
```

```
        case when (
                select distinct brcid
                from
SQLCRIS_User.RPatel.rp_firstreferral_psychosis_teamwardbefore_2007_2012
                where (
                        Location_Name like '%OASIS%')
                        and
                        brcid=r.brcid
                ) IS not null then 1 else 0 end PriorOASIS,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(d,14,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
mhasection2wFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,1,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
mhasection1mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,3,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
mhasection3mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,6,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
mhasection6mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
```

```
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,12,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
mhasection12mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,24,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
mhasection24mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.inpatient_episode
                where
                        admission_Date between r.Dateofclosestteamepisode
and
                        DATEADD(d,14,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
inpatient2wFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.inpatient_episode
                where
                        admission_Date between r.Dateofclosestteamepisode
and
                        DATEADD(m,1,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
inpatient1mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.inpatient_episode
                where
                        admission_Date between r.Dateofclosestteamepisode
and
                        DATEADD(m,3,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
inpatient3mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.inpatient_episode
                where
```

```
                        admission_Date between r.Dateofclosestteamepisode
and
                        DATEADD(m,6,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
inpatient6mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.inpatient_episode
                where
                        admission_Date between r.Dateofclosestteamepisode
and
                        DATEADD(m,12,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
inpatient12mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from sqlcris.dbo.inpatient_episode
                where
                        admission_Date between r.Dateofclosestteamepisode
and
                        DATEADD(m,24,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
inpatient24mFirstPsychosisReferral,
        case when (
                select distinct brcid
                from
SQLCRIS_User.RPatel.rp_medication_combined_antipsychotic_recode
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(d,14,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
antipsychotic2wFirstPsychosisReferral,
        case when (
                select distinct brcid
                from
SQLCRIS_User.RPatel.rp_medication_combined_antipsychotic_recode
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,1,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
antipsychotic1mFirstPsychosisReferral,

        case when (
                select distinct brcid
```

```
                from
SQLCRIS_User.RPatel.rp_medication_combined_antipsychotic_recode
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,3,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
antipsychotic3mFirstPsychosisReferral,

        case when (
                select distinct brcid
                from
SQLCRIS_User.RPatel.rp_medication_combined_antipsychotic_recode
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,6,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
antipsychotic6mFirstPsychosisReferral,

        case when (
                select distinct brcid
                from
SQLCRIS_User.RPatel.rp_medication_combined_antipsychotic_recode
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,12,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
antipsychotic12mFirstPsychosisReferral,

        case when (
                select distinct brcid
                from
SQLCRIS_User.RPatel.rp_medication_combined_antipsychotic_recode
                where
                        Start_Date between r.Dateofclosestteamepisode and
                        DATEADD(m,24,r.Dateofclosestteamepisode) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end
antipsychotic24mFirstPsychosisReferral,
        (SELECT           COUNT(*) AS noofadmissions
                FROM  sqlcris.dbo.inpatient_episode
                WHERE (BrcId = r.brcid) AND
                (Admission_Date <= DATEADD(w,2,r.Dateofclosestteamepisode))
AND
                (Discharge_Date >= r.Dateofclosestteamepisode) OR
                (BrcId = r.brcid) AND
```

```
                     (Admission_Date <= DATEADD(w,2,r.Dateofclosestteamepisode))
AND
                     (Discharge_Date = '1 Jan 1900')
                     GROUP BY BrcId) AS admissioncount2w,
          (SELECT             COUNT(*) AS noofadmissions
                     FROM  sqlcris.dbo.inpatient_episode
                     WHERE (BrcId = r.brcid) AND
                     (Admission_Date <= DATEADD(m,1,r.Dateofclosestteamepisode))
AND
                     (Discharge_Date >= r.Dateofclosestteamepisode) OR
                     (BrcId = r.brcid) AND
                     (Admission_Date <= DATEADD(m,1,r.Dateofclosestteamepisode))
AND
                     (Discharge_Date = '1 Jan 1900')
                     GROUP BY BrcId) AS admissioncount1m,
          (SELECT             COUNT(*) AS noofadmissions
                     FROM  sqlcris.dbo.inpatient_episode
                     WHERE (BrcId = r.brcid) AND
                     (Admission_Date <= DATEADD(m,3,r.Dateofclosestteamepisode))
AND
                     (Discharge_Date >= r.Dateofclosestteamepisode) OR
                     (BrcId = r.brcid) AND
                     (Admission_Date <= DATEADD(m,3,r.Dateofclosestteamepisode))
AND
                     (Discharge_Date = '1 Jan 1900')
                     GROUP BY BrcId) AS admissioncount3m,
          (SELECT             COUNT(*) AS noofadmissions
                     FROM  sqlcris.dbo.inpatient_episode
                     WHERE (BrcId = r.brcid) AND
                     (Admission_Date <= DATEADD(m,6,r.Dateofclosestteamepisode))
AND
                     (Discharge_Date >= r.Dateofclosestteamepisode) OR
                     (BrcId = r.brcid) AND
                     (Admission_Date <= DATEADD(m,6,r.Dateofclosestteamepisode))
AND
                     (Discharge_Date = '1 Jan 1900')
                     GROUP BY BrcId) AS admissioncount6m,
          (SELECT             COUNT(*) AS noofadmissions
                     FROM  sqlcris.dbo.inpatient_episode
                     WHERE (BrcId = r.brcid) AND
                     (Admission_Date <=
DATEADD(m,12,r.Dateofclosestteamepisode)) AND
                     (Discharge_Date >= r.Dateofclosestteamepisode) OR
                     (BrcId = r.brcid) AND
                     (Admission_Date <=
DATEADD(m,12,r.Dateofclosestteamepisode)) AND
                     (Discharge_Date = '1 Jan 1900')
                     GROUP BY BrcId) AS admissioncount12m,
          (SELECT             COUNT(*) AS noofadmissions
```

```
                FROM  sqlcris.dbo.inpatient_episode
                WHERE (BrcId = r.brcid) AND
                (Admission_Date <=
DATEADD(m,24,r.Dateofclosestteamepisode)) AND
                (Discharge_Date >= r.Dateofclosestteamepisode) OR
                (BrcId = r.brcid) AND
                (Admission_Date <=
DATEADD(m,24,r.Dateofclosestteamepisode)) AND
                (Discharge_Date = '1 Jan 1900')
                GROUP BY BrcId) AS admissioncount24m,
        SQLCrisImport.dbo.getlos(Dateofclosestteamepisode, DATEADD(m, 12,
Dateofclosestteamepisode), r.brcid) AS los_PsychosisReferralDate12m,
        SQLCrisImport.dbo.getlos(Dateofclosestteamepisode, DATEADD(m, 24,
Dateofclosestteamepisode), r.brcid) AS los_PsychosisReferralDate24m
from [SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_recode_2007_2012]
r
left join sqlcris.dbo.epr_form p on r.brcid=p.brcid
left join [SQLCrisImport].[dbo].[address_ons_2010_imd] h on h.CN_Doc_ID=
(select top 1 CN_Doc_ID from [SQLCrisImport].[dbo].[address_ons_2010_imd]
h1 where h1.BrcId=r.BrcId order by h1.Start_Date asc)
```

In this script, a series of table joins with the EPR form table ("[sqlcris].[dbo].[epr_form]") and a view which stores address data ("[SQLCrisImport].[dbo].[address_ons_2010_imd]") were used to obtain demographic data (age, gender, ethnicity, borough of residence, marital status, employment status and accommodation status). CASE WHEN statements containing sub-queries were used to obtain data on whether patients were diagnosed with a psychotic disorder in a community first-episode psychosis Early Intervention Service ("PriorEISCommunity") represented by the Lambeth Early Onset community and crisis services (LEO), the Southwark Team for Early Intervention in Psychosis (STEP), the Lewisham Early Intervention Service and the Croydon Outreach Assessment Support Team (COAST). Further CASE WHEN statements with sub-queries identified patients presenting to the LEO inpatient service (LEO Unit) and the high risk clinical service (OASIS). The latter was used as an exclusion filter in order to prevent overlap of the sample extracted using the BRC Case Register and the OASIS sample manually extracted using ePJS. The CASE WHEN statements identifying whether patients had presented to EIS Community, EIS Inpatient or the OASIS service drew data from the

"[SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_teamwardbefore_2007_2012]" view, which is reproduced and described further in section 3.22.

The remaining CASE WHEN statements in the prior query included sub-queries to extract data on clinical outcomes at various time points from 2 weeks to 2 years. These included compulsory admission under the UK Mental Health Act ("mhasectionXXFirstPsychosisReferral"), any inpatient admission ("inpatientXXFirstPsychosisReferral"), whether started on antipsychotic medication ("antipsychoticXXFirstPsychosisReferral") and the number of hospital admissions ("admissioncountXX") during the follow-up period. The follow-up period was taken from the date of being accepted to a clinical service during which the diagnosis of a psychotic disorder was made ("Dateofclosestteamepisode"). The final two outputs in the SELECT statement initiated a SQL store procedure ("SQLCrisImport.dbo.getlos") designed to identify the number of days during a given period in which a patient was admitted to hospital ("los_PsychosisReferralDateXXX").

The output of this query was saved as a view in the following location: "[SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_recode_2007_2012output]" from which the next script was applied.

```
USE [SQLCRIS_User]
SELECT [brcid]
       ,[Dateoffirstreferral]
       ,YEAR([Dateoffirstreferral]) Yearoffirstreferral
       ,[Dateofclosestteamepisode]
       ,[Dateoffirstpsychosisdiagnosis]
       ,YEAR([Dateoffirstpsychosisdiagnosis]) Yearoffirstpsychosisdiagnosis
       ,[Firstpsychosisdiagnosisraw]
       ,[Firstpsychosisdiagnosisrecode]
         ,case
                when (
                         [Firstpsychosisdiagnosisrecode] like
'F2xSchizophrenia')
                then '1'

                when (
```

```
                        [Firstpsychosisdiagnosisrecode] like 'Bipolar')
                then '2'

                when (
                        [Firstpsychosisdiagnosisrecode] like
'PsychoticDepression')
                then '3'

                when (
                        [Firstpsychosisdiagnosisrecode] like
'Schizoaffective')
                then '4'

                when (
                        [Firstpsychosisdiagnosisrecode] like
'F1x.5DrugPsyc')
                then '5'

                when (
                        [Firstpsychosisdiagnosisrecode] like
'OtherPsychosis')
                then '6'

                end as Firstpsychosisdiagnosisrecode2

      ,[Borough]
        ,case
                when (
                        Borough like '%Lambeth%')
                then 'Lambeth'

                when (
                        Borough like '%Croydon%')
                then 'Croydon'

                when (
                        Borough like '%Southwark%')
                then 'Southwark'

                when (
                        Borough like '%Lewisham%')
                then 'Lewisham'

                when (
                        Borough is NULL)
                then 'NotRecorded'

                else 'Other'
```

```
            end as BoroughRecode,
    case
            when (
                    Borough like '%Lambeth%')
            then '2'

            when (
                    Borough like '%Croydon%')
            then '3'

            when (
                    Borough like '%Southwark%')
            then '4'

            when (
                    Borough like '%Lewisham%')
            then '5'

            when (
                    Borough is NULL)
            then '6'

            else '1'

            end as BoroughRecode2
    ,[DiagnosticDelayInDays]
    ,[age]
    ,[Gender]
    ,case
            when
                    Gender like 'Female'
            then '0'

            when
                    Gender like 'Male'
            then '1'

            when
                    Gender like 'Not Known'
            then '2'

            end as GenderRecode2
    ,[ethnicity]
    ,case
            when (
                    ethnicity like 'Irish (B)' or
                    ethnicity like 'Any other white background (C)' or
                    ethnicity like 'British (A)')
            then 'White'
```

```
            when (
                    ethnicity like 'Bangladeshi (K)' or
                    ethnicity like 'Pakistani (J)' or
                    ethnicity like 'Chinese (R)' or
                    ethnicity like 'Indian (H)' or
                    ethnicity like 'Any other Asian background (L)')
            then 'Asian'

            when (
                    ethnicity like 'African (N)')
            then 'BlackAfrican'

            when (
                    ethnicity like 'Caribbean (M)')
            then 'BlackCaribbean'

            when (
                    ethnicity like 'Any other black background (P)')
            then 'BlackOther'

            when (
                    ethnicity like 'White and Asian (F)' or
                    ethnicity like 'White and Black African (E)' or
                    ethnicity like 'Any other mixed background (G)' or
                    ethnicity like 'Any other ethnic group (S)')
            then 'Other'

            when (
                    ethnicity like 'Not Stated (Z)' or
                    ethnicity like 'None')
            then 'NotRecorded'

            end as ethnicrecode,
    case
            when (
                    ethnicity like 'Irish (B)' or
                    ethnicity like 'Any other white background (C)' or
                    ethnicity like 'British (A)')
            then '1'

            when (
                    ethnicity like 'Bangladeshi (K)' or
                    ethnicity like 'Pakistani (J)' or
                    ethnicity like 'Chinese (R)' or
                    ethnicity like 'Indian (H)' or
                    ethnicity like 'Any other Asian background (L)')
            then '2'
```

```
                when (
                        ethnicity like 'African (N)')
                then '3'

                when (
                        ethnicity like 'Caribbean (M)')
                then '4'

                when (
                        ethnicity like 'Any other black background (P)')
                then '5'

                when (
                        ethnicity like 'White and Asian (F)' or
                        ethnicity like 'White and Black African (E)' or
                        ethnicity like 'Any other mixed background (G)' or
                        ethnicity like 'Any other ethnic group (S)')
                then '6'

                when (
                        ethnicity like 'Not Stated (Z)' or
                        ethnicity like 'None')
                then '7'

                end as ethnicrecode2
        ,[marital_status]
        ,case
                when (
                        marital_status like 'Cohabiting' or
                        marital_status like 'Married' or
                        marital_status like 'Married/Civil Partner')
                then 'MarriedCohabiting'

                when (
                        marital_status like 'Divorced' or
                        marital_status like 'Divorced/Civil Partnership
Dissolved' or

                        marital_status like 'Separated')
                then 'DivorcedSeparated'

                when (
                        marital_status like 'Single')
                then 'Single'

                when (
                        marital_status like 'Widowed' or
                        marital_status like 'Widowed/Surviving Civil
Partner')
                then 'Widowed'
```

```
                when (
                        marital_status like 'Not Disclosed' or
                        marital_status like 'Not Known')
                then 'NotRecorded'

                end as maritalrecode,
      case
                when (
                        marital_status like 'Cohabiting' or
                        marital_status like 'Married' or
                        marital_status like 'Married/Civil Partner')
                then '1'

                when (
                        marital_status like 'Divorced' or
                        marital_status like 'Divorced/Civil Partnership
Dissolved' or

                        marital_status like 'Separated')
                then '2'

                when (
                        marital_status like 'Single')
                then '3'

                when (
                        marital_status like 'Widowed' or
                        marital_status like 'Widowed/Surviving Civil
Partner')
                then '4'

                when (
                        marital_status like 'Not Disclosed' or
                        marital_status like 'Not Known')
                then '5'

                end as maritalrecode2
      ,[Employment_status]
      ,case
                when (
                        Employment_status like 'Volunteer' or
                        Employment_status like 'Self Employed' or
                        Employment_status like 'Part Time Employment' or
                        Employment_status like 'Paid Employment')
                then 'Employed'

                when (
                        Employment_status like 'Govt Training Scheme' or
                        Employment_status like 'Full Time Student' or
```

```
                Employment_status like 'Full Time Student - School
age')
            then 'Student'

            when (
                Employment_status like 'Retired')
            then 'Retired'

            when (
                Employment_status like 'Registered Disabled' or
                Employment_status like 'Unemployed')
            then 'Unemployed'

            when (
                Employment_status like 'Other' or
                Employment_status like 'Not Known' or
                Employment_status like 'xNx')
            then 'NotRecorded'

            end as employmentrecode,
      case
            when (
                Employment_status like 'Volunteer' or
                Employment_status like 'Self Employed' or
                Employment_status like 'Part Time Employment' or
                Employment_status like 'Paid Employment')
            then '1'

            when (
                Employment_status like 'Govt Training Scheme' or
                Employment_status like 'Full Time Student' or
                Employment_status like 'Full Time Student - School
age')
            then '2'

            when (
                Employment_status like 'Retired')
            then '3'

            when (
                Employment_status like 'Registered Disabled' or
                Employment_status like 'Unemployed')
            then '4'

            when (
                Employment_status like 'Other' or
                Employment_status like 'Not Known' or
                Employment_status like 'xNx')
            then '5'
```

```
            end as employmentrecode2
      ,[Accommodation_status]
      ,case
            when (
                  Accommodation_status like 'Owner')
            then 'Owner'

            when (
                  Accommodation_status like 'Private Tenant')
            then 'PrivateTenant'

            when (
                  Accommodation_status like 'Council Tenant')
            then 'CouncilTenant'

            when (
                  Accommodation_status like 'Trust' or
                  Accommodation_status like 'Nursing/Residential')
            then 'SupportedAccommodation'

            when (
                  Accommodation_status like 'Homeless')
            then 'Homeless'

            when (
                  Accommodation_status like 'Other')
            then 'Other'

            when (
                  Accommodation_status like 'Not known' or
                  Accommodation_status like 'xNx')
            then 'NotRecorded'

            end as accommodationrecode,
      case
            when (
                  Accommodation_status like 'Owner')
            then '1'

            when (
                  Accommodation_status like 'Private Tenant')
            then '2'

            when (
                  Accommodation_status like 'Council Tenant')
            then '3'

            when (
```

```
                        Accommodation_status like 'Trust' or
                        Accommodation_status like 'Nursing/Residential')
                then '4'

                when (
                        Accommodation_status like 'Homeless')
                then '5'

                when (
                        Accommodation_status like 'Other')
                then '6'

                when (
                        Accommodation_status like 'Not known' or
                        Accommodation_status like 'xNx')
                then '7'

                end as accommodationrecode2
        ,[PriorEISCommunity]
        ,[PriorEISInpatientLEO]
          ,case
                when (
                        PriorEISCommunity=1 and
                        PriorEISInpatientLEO=1)
                then '1' else '0'
                end as PriorEIS
        ,[PriorOASIS]
        ,[mhasection2wFirstPsychosisReferral]
        ,[mhasection1mFirstPsychosisReferral]
        ,[mhasection3mFirstPsychosisReferral]
        ,[mhasection6mFirstPsychosisReferral]
        ,[mhasection12mFirstPsychosisReferral]
        ,[mhasection24mFirstPsychosisReferral]
        ,[inpatient2wFirstPsychosisReferral]
        ,[inpatient1mFirstPsychosisReferral]
        ,[inpatient3mFirstPsychosisReferral]
        ,[inpatient6mFirstPsychosisReferral]
        ,[inpatient12mFirstPsychosisReferral]
        ,[inpatient24mFirstPsychosisReferral]
        ,[antipsychotic2wFirstPsychosisReferral]
        ,[antipsychotic1mFirstPsychosisReferral]
        ,[antipsychotic3mFirstPsychosisReferral]
        ,[antipsychotic6mFirstPsychosisReferral]
        ,[antipsychotic12mFirstPsychosisReferral]
        ,[antipsychotic24mFirstPsychosisReferral]
        ,[admissioncount2w]
        ,[admissioncount1m]
        ,[admissioncount3m]
        ,[admissioncount6m]
```

```
      ,[admissioncount12m]
      ,[admissioncount24m]
      ,[los_PsychosisReferralDate12m]
      ,[los_PsychosisReferralDate24m]
  FROM
[SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_recode_2007_2012output]
where (age between 14 and 35) and PriorOASIS=0
```

In this script, a series of CASE WHEN statements was used to recode predictor variables in preparation for statistical analysis. A WHERE clause was used to filter the sample to only include patients between the ages of 14 and 35 and who had not presented with psychosis to the high risk clinical service (“PriorOASIS=0”).

The output of this query was exported into a Microsoft Excel spread sheet and imported into STATA where it was merged with data from another spread sheet containing the manually extracted data on patients presenting to the high risk clinical service. The merged dataset was analysed using the methods described in the publication in section 3.1.

3.22 SQL support queries

The following query was saved as a view (“[SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_teamwardbefore_2007_2012]”) to facilitate queries relating to which teams patients had presented to prior to their psychotic disorder diagnosis.

```
USE [SQLCRIS_User]
SELECT v.brcid, 'Team' [type], te.Accepted_Date start_date,
te.discharge_date end_date, Location_Name, cag
        FROM [SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_2007_2012]
v LEFT JOIN
        sqlcris.dbo.Team_episode te ON v.brcid = te.BrcId
WHERE      (te.accepted_Date > '1-jan-1900' AND te.Rejection_Date = '1-jan-
1900') AND te.Location_Name NOT LIKE '%test%' AND
                      (te.Accepted_Date <= v.Dateoffirstpsychosisdiagnosis
AND te.Accepted_Date > '1-jan-1900' AND (te.Discharge_Date >=
v.accepted_date OR
                      te.Discharge_Date = '01-jan-1900'))
```

```
/*----------------ward*/ UNION
SELECT     v.brcid, 'Ward' [type], Actual_Start_Date start_date,
Actual_End_Date end_date, Location_Name, cag
FROM          [SQLCRIS_User].[RPatel].[rp_firstreferral_psychosis_2007_2012]
v LEFT JOIN
                      sqlcris.dbo.Ward_stay te ON v.brcid = te.BrcId
WHERE      ((te.Actual_Start_Date > '1-jan-1900' AND te.Rejection_Date = '1-
jan-1900') AND te.Location_Name NOT LIKE '%test%' AND
                      (te.Current_Ward_Stay_Status_ID LIKE 'closed' OR
                      te.Current_Ward_Stay_Status_ID LIKE '%occupied%'))
AND (Actual_Start_Date <= v.Dateoffirstpsychosisdiagnosis AND
                      Actual_Start_Date > '1-jan-1900' AND (Actual_End_Date
>= v.accepted_date OR
                      Actual_End_Date = '01-jan-1900'))
```

This view joined data of all team episodes and hospital ward admissions prior to the date of psychosis diagnosis (“v.Dateoffirstpsychosisdiagnosis”) into a single view using sub-queries on the “[sqlcris].[dbo].[Team_episode]” and “[sqlcris].[dbo].[Ward_stay]” tables which are joined together using a UNION statement. In this way, all community and inpatient episodes for patients in the study prior to their psychotic disorder diagnoses were represented in a single view in order to avoid having to search each of the sources separately in the main SQL queries.

3.23 Statistical analysis

There were five permutations of statistical analysis performed in this study. The STATA do files employed to perform each analysis are reproduced below. Each do file contained a common header used to perform the following actions:

(i) Remove participants with no recorded gender. This was a data quality procedure as all such participants did not have any meaningful covariate data.

(ii) Change the variable type of the Diagnostic Delay variable from text string to numerical.

(iii) Recode ethnicity into a smaller number of categories by placing “Black African”, “Black Caribbean” and “Black Other” into a single category.

(iv) Recode new variables for covariates to take into account missing data (specified in STATA by “.”).

(v) Generate a binary variable called “fullcovariate” to indicate any participants with missing data.

(vi) Recode the “PriorEIS” variable into a three category variable specifying whether patients presented to the high risk clinical service (OASIS) a first episode early intervention service (EIS) or another conventional mental health service.

(vii) Generate a variable called “ePJS” to indicate whether or not patients presenting to the OASIS high risk clinical service did so prior to or after the implementation of ePJS, in order to facilitate a sensitivity analysis to see if there were any significant demographic differences between these two groups of patients.

(viii) Add category labels to variables.

The do file specifying the common header is reproduced below:

```
**Drop one case with no recorded gender**
drop if GenderRecode2==2

**Destring diagnostic delay**
destring DiagnosticDelayInDays, replace
recode DiagnosticDelayInDays .=0

**Recoding demographic variables to deal with missing data and regrouping
categories**
gen ethnicrecode3 = ethnicrecode2
recode ethnicrecode3 1=1 2=2 3=3 4=3 5=3 6=4 7=5
gen ethnicrecode3b = ethnicrecode3
recode ethnicrecode3b 5=.
gen maritalrecode3 = maritalrecode2
recode maritalrecode3 1=1 2=2 3=3 4=3 5=4
gen maritalrecode3b = maritalrecode3
recode maritalrecode3b 4=.
gen employmentrecode3 = employmentrecode2
recode employmentrecode3 1=1 2=2 3=3 4=3 5=4
gen employmentrecode3b = employmentrecode3
recode employmentrecode3b 4=.
gen accommodationrecode3 = accommodationrecode2
recode accommodationrecode3 7=6
gen accommodationrecode3b = accommodationrecode2
```

```
recode accommodationrecode3b 7=.
gen BoroughRecode3b = BoroughRecode2
recode BoroughRecode3b 5=.
gen fullcovariate = 1
replace fullcovariate = 0 if
ethnicrecode3b==.|maritalrecode3b==.|employmentrecode3b==.|BoroughRecode3b=
=.
replace PriorEIS = 2 if PriorOASIS==1

**Adding variable to test difference between 2001-2006 and 2007-2011**
gen ePJS = Yearoffirstreferral
recode ePJS 2001/2006=0 2007/max=1

**Label variables and create ordinal exposures**
label define Firstpsychosisdiagnosisrecode2 1 "F2xSchizophrenia" 2
"Bipolar" 3 "PsychoticDepression" 4 "Schizoaffective" 5 "F1x.5DrugPsyc" 6
"OtherPsychosis"
label define AgeRecode2 1 "<16" 2 "16-24" 3 "24-49" 4 "50-64" 5 ">64"
label define GenderRecode2 0 "Female" 1 "Male" 2 "NotRecorded"
label define ethnicrecode2 1 "White" 2 "Asian" 3 "BlackAfrican" 4
"BlackCaribbean" 5 "BlackOther" 6 "Other" 7 "NotRecorded"
label define ethnicrecode3 1 "White" 2 "Asian" 3 "Black" 4 "Other" 5
"NotRecorded"
label define ethnicrecode3b 1 "White" 2 "Asian" 3 "Black" 4 "Other"
label define maritalrecode2 1 "MarriedCohabiting" 2 "DivorcedSeparated" 3
"Single" 4 "Widowed" 5 "NotRecorded"
label define maritalrecode3 1 "MarriedCohabiting" 2 "DivorcedSeparated" 3
"Single" 4 "NotRecorded"
label define maritalrecode3b 1 "MarriedCohabiting" 2 "DivorcedSeparated" 3
"Single"
label define employmentrecode2 1 "Employed" 2 "Student" 3 "Retired" 4
"Unemployed" 5 "NotRecorded"
label define employmentrecode3 1 "Employed" 2 "Student" 3 "Unemployed" 4
"NotRecorded"
label define employmentrecode3b 1 "Employed" 2 "Student" 3 "Unemployed"
label define accommodationrecode2 1 "Owner" 2 "PrivateTenant" 3
"CouncilTenant" 4 "SupportedAccomodation" 5 "Homeless" 6 "Other" 7
"NotRecorded"
label define accommodationrecode3 1 "Owner" 2 "PrivateTenant" 3
"CouncilTenant" 4 "SupportedAccomodation" 5 "Homeless" 6 "Other"
label define accommodationrecode3b 1 "Owner" 2 "PrivateTenant" 3
"CouncilTenant" 4 "SupportedAccomodation" 5 "Homeless" 6 "Other"
label define PriorEIS 0 "StandardMentalHealth" 1 "EIS" 2 "OASIS"
label define BoroughRecode2 1 "Lambeth" 2 "Southwark" 3 "Lewisham" 4
"Croydon" 5 "Other"
label define BoroughRecode3b 1 "Lambeth" 2 "Southwark" 3 "Lewisham" 4
"Croydon"
label values Firstpsychosisdiagnosisrecode2 Firstpsychosisdiagnosisrecode2
label values AgeRecode2 AgeRecode2
```

```
label values GenderRecode2 GenderRecode2
label values ethnicrecode2 ethnicrecode2
label values ethnicrecode3 ethnicrecode3
label values ethnicrecode3b ethnicrecode3b
label values maritalrecode2 maritalrecode2
label values maritalrecode3 maritalrecode3
label values maritalrecode3b maritalrecode3b
label values employmentrecode2 employmentrecode2
label values employmentrecode3 employmentrecode3
label values employmentrecode3b employmentrecode3b
label values accommodationrecode2 accommodationrecode2
label values accommodationrecode3 accommodationrecode3
label values accommodationrecode3b accommodationrecode3b
label values PriorEIS PriorEIS
label values BoroughRecode2 BoroughRecode2
label values BoroughRecode3b BoroughRecode3b

**N.B. all follow up period taken from referral date**
```

For each of the five analyses, the following STATA commands were appended to the common header as follows:

(i) Descriptive statistics comparing high risk clinical service and conventional mental health services (presented in main manuscript).

```
**Drop if prior to ePJS**
**drop if ePJS==0**

**Drop if after ePJS**
**drop if ePJS==1**

**Test of normality for age**
ksmirnov age, by(PriorOASIS)
**non-normal distribution for age by PriorOASIS - use Mann Whitney U test**

**Check how many have full covariate data**
tab PriorOASIS if fullcovariate==1

**Demographic variables**
**Chi2 and Fisher's exact test**
summ age if PriorOASIS==0, detail
summ age if PriorOASIS==1, detail
```

```
ranksum age, by(PriorOASIS)
tab PriorOASIS Firstpsychosisdiagnosisrecode2, row chi2
**tab PriorOASIS Firstpsychosisdiagnosisrecode2, row exact**
tab PriorOASIS GenderRecode2, row chi2
tab PriorOASIS GenderRecode2, row exact
tab PriorOASIS ethnicrecode3b, row chi2
**tab PriorOASIS ethnicrecode3b, row exact**
tab PriorOASIS maritalrecode3b, row chi2
**tab PriorOASIS maritalrecode3b, row exact**
tab PriorOASIS employmentrecode3b, row chi2
**tab PriorOASIS employmentrecode3b, row exact**
tab PriorOASIS BoroughRecode3b, row chi2
tab PriorOASIS BoroughRecode3b, row exact

**Demographic variables including missing data**
**Chi2 and Fisher's exact test**
tab PriorOASIS ethnicrecode3, row chi2
**tab PriorOASIS ethnicrecode3, row exact**
tab PriorOASIS maritalrecode3, row chi2
**tab PriorOASIS maritalrecode3, row exact**
tab PriorOASIS employmentrecode3, row chi2
**tab PriorOASIS employmentrecode3, row exact**
tab PriorOASIS BoroughRecode2, row chi2
**tab PriorOASIS BoroughRecode2, row exact**

**MHA Section**
tab PriorOASIS mhasection2w, row
tab PriorOASIS mhasection1m, row
tab PriorOASIS mhasection3m, row
tab PriorOASIS mhasection6m, row
tab PriorOASIS mhasection12m, row
tab PriorOASIS mhasection24m, row

**Number of admissions**
summ admissioncount2w if PriorOASIS==0, detail
summ admissioncount2w if PriorOASIS==1, detail
summ admissioncount1m if PriorOASIS==0, detail
summ admissioncount1m if PriorOASIS==1, detail
summ admissioncount3m if PriorOASIS==0, detail
summ admissioncount3m if PriorOASIS==1, detail
summ admissioncount6m if PriorOASIS==0, detail
summ admissioncount6m if PriorOASIS==1, detail
summ admissioncount12m if PriorOASIS==0, detail
summ admissioncount12m if PriorOASIS==1, detail
summ admissioncount24m if PriorOASIS==0, detail
summ admissioncount24m if PriorOASIS==1, detail

**Number of inpatient days**
summ los_12m if PriorOASIS==0, detail
```

```
summ los_12m if PriorOASIS==1, detail
summ los_24m if PriorOASIS==0, detail
summ los_24m if PriorOASIS==1, detail

**Diagnostic delay**
summ DiagnosticDelayInDays, detail
summ DiagnosticDelayInDays if PriorOASIS==0, detail
summ DiagnosticDelayInDays if PriorOASIS==1, detail

/*
graph box DiagnosticDelayInDays, over(PriorOASIS, relabel(1 "Conventional
mental health service" 2 "High risk service") descending) box(1,
fcolor(navy) fintensity(inten100) lcolor(black) lpattern(solid)) box(2,
fcolor(cranberry) fintensity(inten100) lcolor(black) lpattern(solid))
ytitle(Time to diagnosis (days)) ytitle(, size(small)) title(Time to
diagnosis in high risk service compared to conventional mental health
services, size(medsmall)) nooutsides
*/
```

This do file estimated descriptive statistics by the PriorOASIS variable which indicated whether or not patients presented to the high risk clinical service. The additional commands commented out at the beginning were uncommented sequentially to provide data for the sensitivity analysis to compare between patients presenting prior to and after the implementation of ePJS. The command commented out at the end generated a box plot summarising differences in diagnostic delay between the groups. This was presented as Figure 1 in the main manuscript (section 3.1).

(ii) Descriptive statistics comparing high risk clinical service, first episode services and other conventional mental health services (presented in supplementary material).

```
**Check how many have full covariate data**
tab PriorEIS if fullcovariate==1

**Demographic variables**
**Chi2 and Fisher's exact test**
summ age if PriorEIS==0, detail
summ age if PriorEIS==1, detail
```

```
summ age if PriorEIS==2, detail
anova age PriorEIS
tab PriorEIS Firstpsychosisdiagnosisrecode2, row chi2
**tab PriorEIS Firstpsychosisdiagnosisrecode2, row exact**
tab PriorEIS GenderRecode2, row chi2
tab PriorEIS GenderRecode2, row exact
tab PriorEIS ethnicrecode3b, row chi2
**tab PriorEIS ethnicrecode3b, row exact**
tab PriorEIS maritalrecode3b, row chi2
**tab PriorEIS maritalrecode3b, row exact**
tab PriorEIS employmentrecode3b, row chi2
**tab PriorEIS employmentrecode3b, row exact**
tab PriorEIS BoroughRecode3b, row chi2
**tab PriorEIS BoroughRecode3b, row exact**

**Demographic variables including missing data**
**Chi2 and Fisher's exact test**
tab PriorEIS ethnicrecode3, row chi2
**tab PriorEIS ethnicrecode3, row exact**
tab PriorEIS maritalrecode3, row chi2
**tab PriorEIS maritalrecode3, row exact**
tab PriorEIS employmentrecode3, row chi2
**tab PriorEIS employmentrecode3, row exact**
tab PriorEIS BoroughRecode2, row chi2
**tab PriorEIS BoroughRecode2, row exact**

**MHA Section**
tab PriorEIS mhasection2w, row
tab PriorEIS mhasection1m, row
tab PriorEIS mhasection3m, row
tab PriorEIS mhasection6m, row
tab PriorEIS mhasection12m, row
tab PriorEIS mhasection24m, row

**Number of admissions**
summ admissioncount2w if PriorEIS==0, detail
summ admissioncount2w if PriorEIS==1, detail
summ admissioncount2w if PriorEIS==2, detail
summ admissioncount1m if PriorEIS==0, detail
summ admissioncount1m if PriorEIS==1, detail
summ admissioncount1m if PriorEIS==2, detail
summ admissioncount3m if PriorEIS==0, detail
summ admissioncount3m if PriorEIS==1, detail
summ admissioncount3m if PriorEIS==2, detail
summ admissioncount6m if PriorEIS==0, detail
summ admissioncount6m if PriorEIS==1, detail
summ admissioncount6m if PriorEIS==2, detail
summ admissioncount12m if PriorEIS==0, detail
summ admissioncount12m if PriorEIS==1, detail
```

```
summ admissioncount12m if PriorEIS==2, detail
summ admissioncount24m if PriorEIS==0, detail
summ admissioncount24m if PriorEIS==1, detail
summ admissioncount24m if PriorEIS==2, detail

**Number of inpatient days**
summ los_12m if PriorEIS==0, detail
summ los_12m if PriorEIS==1, detail
summ los_12m if PriorEIS==2, detail
summ los_24m if PriorEIS==0, detail
summ los_24m if PriorEIS==1, detail
summ los_24m if PriorEIS==2, detail

**Diagnostic delay**
summ DiagnosticDelayInDays, detail
summ DiagnosticDelayInDays if PriorEIS==0, detail
summ DiagnosticDelayInDays if PriorEIS==1, detail
summ DiagnosticDelayInDays if PriorEIS==2, detail
```

This do file followed the same principles as (i) but instead separating by the PriorEIS variable indicating whether patients presented to the high risk clinical service, a first episode early intervention service or another conventional mental health service. The results were presented in the supplementary material eTable 4 (section 3.1).

(iii) Regression analyses including missing covariate data comparing high risk clinical service and conventional mental health services (presented in main manuscript).

```
**Multivariable analyses with missing data included**

**MHA section**
logistic mhasection2w i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic2w
logistic mhasection1m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic1m
```

```
logistic mhasection3m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic3m
logistic mhasection6m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic6m
logistic mhasection12m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic12m
logistic mhasection24m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m

**Admission count**
poisson admissioncount2w i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic2w, irr
poisson admissioncount1m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic1m, irr
poisson admissioncount3m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic3m, irr
poisson admissioncount6m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic6m, irr
poisson admissioncount12m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic12m, irr
poisson admissioncount24m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m, irr

**Inpatient days from accepted date**
regress los_12m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic12m
regress los_24m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m

**Diagnostic delay**
regress DiagnosticDelayInDays i.PriorOASIS i.Firstpsychosisdiagnosisrecode2
age i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m
```

The output of these commands produced the results of multivariable regression analysis investigating the association of presentation to the high risk clinical service with compulsory admission under the UK Mental Health Act (binary logistic regression), number of days spent in hospital (Poisson regression), number of days spent in hospital and diagnostic delay (multiple linear regression). All analyses were adjusted for diagnosis, age, gender, ethnicity, marital status, employment status, borough of residence and whether started on antipsychotic medication. These results were presented as Table 2a and 2b in the main manuscript (section 3.1).

(iv) Regression analyses including missing covariate data comparing high risk clinical service, first episode services and other conventional mental health services (presented in supplementary material).

```
**Multivariable analyses with missing data included**

**MHA section**
logistic mhasection2w i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic2w
logistic mhasection1m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic1m
logistic mhasection3m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic3m
logistic mhasection6m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic6m
logistic mhasection12m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic12m
logistic mhasection24m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m

**Admission count**
```

```
poisson admissioncount2w i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic2w, irr
poisson admissioncount1m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic1m, irr
poisson admissioncount3m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic3m, irr
poisson admissioncount6m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic6m, irr
poisson admissioncount12m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic12m, irr
poisson admissioncount24m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m, irr

**Inpatient days from accepted date**
regress los_12m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic12m
regress los_24m i.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m

**Diagnostic delay**
regress DiagnosticDelayInDays i.PriorEIS i.Firstpsychosisdiagnosisrecode2
age i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m

**Multivariable analyses with missing data included - EIS as reference**

**MHA section**
logistic mhasection2w ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic2w
logistic mhasection1m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic1m
logistic mhasection3m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic3m
logistic mhasection6m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic6m
```

```
logistic mhasection12m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic12m
logistic mhasection24m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m

**Admission count**
poisson admissioncount2w ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic2w, irr
poisson admissioncount1m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic1m, irr
poisson admissioncount3m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic3m, irr
poisson admissioncount6m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic6m, irr
poisson admissioncount12m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic12m, irr
poisson admissioncount24m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m, irr

**Inpatient days from accepted date**
regress los_12m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic12m
regress los_24m ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m

**Diagnostic delay**
regress DiagnosticDelayInDays ib1.PriorEIS i.Firstpsychosisdiagnosisrecode2
age i.GenderRecode2 i.ethnicrecode3 i.maritalrecode3 i.employmentrecode3
i.BoroughRecode2 i.antipsychotic24m
```

These commands replaced the PriorOASIS variable with PriorEIS as the predictor variable in the same multivariable regression analyses as in (iii). As PriorEIS contained three categories (high risk clinical service, first episode early intervention services and any other conventional mental health service), two sets of regression analyses were performed. One with any other conventional mental health service as the reference category (i.PriorEIS) and one with first episode early intervention services as the reference category (ib1.PriorEIS). The results of this analysis were presented in eTables 5a/b/c and 6a/b/c in the supplementary material (section 3.1).

(v) Regression analysis including only participants with full covariate data comparing high risk clinical service and conventional mental health services (presented in supplementary material).

```
**Multivariable analyses with missing data dropped**

**MHA section**
logistic mhasection2w i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic2w
logistic mhasection1m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic1m
logistic mhasection3m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic3m
logistic mhasection6m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic6m
logistic mhasection12m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic12m
logistic mhasection24m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic24m

**Admission count**
poisson admissioncount2w i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic2w, irr
poisson admissioncount1m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic1m, irr
```

```
poisson admissioncount3m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic3m, irr
poisson admissioncount6m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic6m, irr
poisson admissioncount12m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic12m, irr
poisson admissioncount24m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic24m, irr

**Inpatient days from accepted date**
regress los_12m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic12m
regress los_24m i.PriorOASIS i.Firstpsychosisdiagnosisrecode2 age
i.GenderRecode2 i.ethnicrecode3b i.maritalrecode3b i.employmentrecode3b
i.BoroughRecode3b i.antipsychotic24m
```

The commands in this analysis are analogous to (iii) but specifying covariates with missing data dropped (i.e. as “.”). The results of this analysis are presented in eTable 3a/b in the supplementary material (section 3.1).

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

This chapter incorporates an article published in BMJ Open which investigates the association of cannabis use with clinical outcomes in FEP, and the possible role of antipsychotic treatment failure in mediating this association. Supplementary methods describing SQL data extraction and statistical analysis using STATA are described in section 4.2.

4.1 BMJ Open journal article

Please see overleaf.

Open Access Research

BMJ Open

Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis: an observational study

Rashmi Patel,[1] Robin Wilson,[1] Richard Jackson,[2] Michael Ball,[2] Hitesh Shetty,[3] Matthew Broadbent,[3] Robert Stewart,[2] Philip McGuire,[1] Sagnik Bhattacharyya[1]

To cite: Patel R, Wilson R, Jackson R, *et al*. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis: an observational study. *BMJ Open* 2016;**6**:e009888. doi:10.1136/bmjopen-2015-009888

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-009888).

PM and SB contributed equally.

Received 1 September 2015
Revised 16 December 2015
Accepted 17 December 2015

For numbered affiliations see end of article.

Correspondence to
Dr Rashmi Patel;
bmj@rpatel.co.uk and
Dr Sagnik Bhattacharyya
sagnik.2.bhattacharyya@kcl.ac.uk

ABSTRACT

Objective: To investigate whether cannabis use is associated with increased risk of relapse, as indexed by number of hospital admissions, and whether antipsychotic treatment failure, as indexed by number of unique antipsychotics prescribed, may mediate this effect in a large data set of patients with first episode psychosis (FEP).

Design: Observational study with exploratory mediation analysis.

Setting: Anonymised electronic mental health record data from the South London and Maudsley NHS Foundation Trust.

Participants: 2026 people presenting to early intervention services with FEP.

Exposure: Cannabis use at presentation, identified using natural language processing.

Main outcome measures: admission to psychiatric hospital and clozapine prescription up to 5 years following presentation.

Mediator: Number of unique antipsychotics prescribed.

Results: Cannabis use was present in 46.3% of the sample at first presentation and was particularly common in patients who were 16–25, male and single. It was associated with increased frequency of hospital admission (incidence rate ratio 1.50, 95% CI 1.25 to 1.80), increased likelihood of compulsory admission (OR 1.55, 1.16 to 2.08) and greater number of days spent in hospital (β coefficient 35.1 days, 12.1 to 58.1). The number of unique antipsychotics prescribed, mediated increased frequency of hospital admission (natural indirect effect 1.09, 95% CI 1.01 to 1.18; total effect 1.50, 1.21 to 1.87), increased likelihood of compulsory admission (natural indirect effect (NIE) 1.27, 1.03 to 1.58; total effect (TE) 1.76, 0.81 to 3.84) and greater number of days spent in hospital (NIE 17.9, 2.4 to 33.4; TE 34.8, 11.6 to 58.1).

Conclusions: Cannabis use in patients with FEP was associated with an increased likelihood of hospital admission. This was linked to the prescription of several different antipsychotic drugs, indicating clinical judgement of antipsychotic treatment failure. Together, this suggests that cannabis use might be associated with worse clinical outcomes in psychosis by contributing towards failure of antipsychotic treatment.

Strengths and limitations of this study

- This is the largest known study (over 2000 participants) to investigate the association of cannabis use with clinical outcome in people with first episode psychosis. As well as demonstrating that cannabis is associated with substantially worse clinical outcomes, our study is the first to identify a possible explanation for these findings through a failure of antipsychotic treatment.
- Our study employed a novel text mining method to identify cannabis use in routinely recorded electronic health record. This approach benefits from increased generalisability of our findings to everyday clinical practice but is limited by the fact that the presence or absence of cannabis use may not have been comprehensively documented in all patients. This may have led to underestimation of its use.
- It was not possible to obtain data on amount, frequency and discontinuation of cannabis use following first presentation to mental health services. Despite this limitation, our data still showed a significant association of cannabis use at presentation to mental health services with poor clinical outcomes up to 5 years later.
- We performed an exploratory mediation analysis to investigate whether the association of cannabis use with poor clinical outcomes could be mediated by an increase in the number of unique antipsychotics prescribed (a marker of antipsychotic treatment failure). However, as this was an observational study, the mediation analysis may have been biased by unmeasured confounders and temporal ambiguity between the mediator and outcome variable.

INTRODUCTION

Cannabis remains the third most common drug of dependence in the world after tobacco and alcohol,[1] with a growing consensus that cannabis use is associated with increased risk of development of psychotic illnesses particularly if used in early adolescence.[2 3] However, there is much less agreement regarding its effect on outcome in those with established psychosis, a substantial proportion of whom use the drug, especially in the early stages of psychosis.[4] Despite the widely held view among clinicians that comorbid cannabis use is a predictor of poor outcome in those with psychosis, the evidence to date has been inconsistent irrespective of the specific outcome measure examined, such as severity of psychotic symptoms or relapse of illness (as indexed by change in symptom severity or hospitalisation), perhaps limited to a large extent by the size of samples and duration of follow-up.[5–10] Since use of cannabis is potentially amenable to treatment, there is a particular need to definitively investigate the effect of comorbid cannabis use on a robust measure of outcome which is indicative of relapse, such as hospitalisation. This is a reliably estimated measure, and has significant implications for the utilisation of healthcare resources.[11]

Furthermore, questions remain as to how cannabis use may increase the risk of relapse. While increased severity of symptoms is likely to play a role, other (but not necessarily unrelated) mechanisms may be through an adverse effect on adherence,[5 12] as well as reduced response to antipsychotic treatment.[13] In a naturalistic setting where decisions regarding medication change take into account a number of factors including response to treatment as well as tolerability and side effects,[14] the number of unique antipsychotics prescribed is a proxy measure which may encompass all these factors. Hence, compared with someone prescribed fewer unique antipsychotics, a person prescribed a greater number of unique antipsychotics may be considered to have a worse antipsychotic response, or in effect, antipsychotic treatment failure, as a result of either treatment resistance or tolerability to the antipsychotic, or a combination of both. However, whether the effect of cannabis use on the increased risk of relapse in psychosis is partly mediated by its effect on antipsychotic treatment failure (as indexed by the number of unique antipsychotics prescribed) has yet to be investigated. Understanding how cannabis use may adversely affect outcome in psychosis is particularly important as it may identify mechanisms that may potentially be amenable to intervention.

In the present study, we attempt to address these issues by investigating the prevalence of cannabis use and its effect on a cohort of patients with first episode psychosis (FEP) receiving mental healthcare from early intervention services. We employed novel data mining and natural language processing (NLP) tools that allowed us to investigate a large data set of anonymised free-text electronic health records in order to obtain data on cannabis use and clinical outcomes. We tested our hypotheses that in those presenting with their first episode of psychosis, cannabis use is associated with increased frequency of hospital admission (including compulsory admission) and greater number of days spent in hospital, and that this is mediated by non-responsiveness to antipsychotics as indexed by the number of unique antipsychotic medications prescribed.

METHODS

Participants

All individuals with FEP who were accepted by an early intervention service in the South London and Maudsley (SLaM) National Health Service (NHS) Foundation Trust between 1 April 2006 and 31 March 2013 were included in the study (n=2026). SLaM is one of the largest providers of specialist mental healthcare in Europe, serving a catchment of around 1.2 million residents in four boroughs of South London (Lambeth, Southwark, Lewisham and Croydon).[15 16]

Source of clinical data

Data for this study were obtained from the SLaM Biomedical Research Centre (BRC) Case Register, which contains anonymised clinical data from the electronic health records of individuals who have previously received or are currently receiving mental healthcare from SLaM.[15] The SLaM BRC Case Register comprises structured fields for demographic information as well as unstructured (but de-identified) free-text fields from case notes and correspondence where history, mental state examination, diagnostic formulation and management plan are primarily recorded. A patient-led oversight committee considers all proposed research before access to the anonymised data is permitted. The electronic health record system was implemented in SLaM early intervention services in April 2006, and so the period of 1 April 2006 to 31 March 2013 was chosen for data capture to maximise the number of participants with at least 1 year of follow-up. Predictor, covariate and outcome variable data were obtained from the SLaM BRC Case Register using the Clinical Record Interactive Search tool (CRIS),[15] a search and database assembly tool underpinning this data resource.[17–20]

Identification of cannabis use

NLP was used to extract documentation of cannabis use from unstructured free-text fields in the BRC Case Register including clinical assessments, reviews and correspondence between healthcare professionals. An NLP application was developed using TextHunter software.[21] Full details of NLP application development are described in a previous study.[22] In summary, a support vector machine learning (SVM) approach was used to identify sentences containing a positive reference of current or historical cannabis use. The application was trained using 478 human-classified sentences which contained the word ‘cannabis’ (or the following synonyms: ‘marijuana’, ‘weed’, ‘pot’, ‘hash’, ‘skunk’, ‘resin’) and optimised using two rounds of active learning classification of a further 1357 sentences. The resulting application was tested against a reference standard of 233 human-classified sentences and an SVM marginal filter applied to obtain a minimum precision value (equivalent to positive predictive value) of 90%. As frequency and amount of cannabis use was not documented in electronic health records in the BRC Case Register, a binary variable defined as any documentation of cannabis use by the patient at presentation with FEP was used. In order to establish baseline cannabis use at the time of presenting with FEP, the cannabis NLP application was applied to clinical records documented within 1 month of presentation to early intervention service as, by this time, all patients would have completed a detailed clinical assessment including assessment of substance use history, allowing a reliable estimation of cannabis exposure at presentation with FEP.

Clinical outcome measures and covariates

The primary outcome was number of psychiatric hospital admissions within the follow-up period. Secondary outcomes included any compulsory hospital admission (under the UK Mental Health Act (MHA)) and number of days spent in hospital during the follow-up period. These outcome measures were obtained from structured fields within the BRC Case Register. The MHA[23] is a UK statute law which allows for compulsory admission to hospital for assessment and/or treatment of a mental illness whose nature and/or degree necessitates hospital admission and where a patient does not consent to be voluntarily admitted. Admission under section 2 of the MHA allows for up to 28 days compulsory admission for assessment of mental illness. Admission under section 3 of the MHA allows for up to 6 months compulsory admission for treatment of mental illness. A patient admitted under section 2 may subsequently be placed under section 3 of the MHA. Compulsory hospital admission in this study was defined as admission to a hospital under section 2 or 3 of the MHA.

The number of unique antipsychotic medications prescribed (as a proxy measure of treatment failure) and whether individuals were prescribed clozapine during the follow-up period were also obtained. The number of unique antipsychotics was analysed as a potential mediating factor in determining association of cannabis use with the primary and secondary outcome variables.

The following variables were extracted as covariates for multivariable analyses: age, gender, ethnicity, marital status and diagnosis. All covariate data obtained were those closest to the date of being accepted by an early intervention service. Ethnicity was recorded according to categories defined by the UK Office for National Statistics.[24] Diagnosis was recorded using the International Classification of Diseases (ICD)-10 classification system, in the following groups: schizophrenia and related disorders (schizophrenia (F20), delusional disorder (F22), schizophrenia-like disorders (F23, F28 and F29)), schizoaffective disorder (F25), mania (F30) or bipolar disorder (F31), psychotic depression (F32.3, F33.3), drug-related psychosis (F1x.5) and other psychotic disorder not otherwise specified. The data were analysed using STATA (V.12) (StataCorp. Stata Statistical Software: Release 12. Coll Station TX StataCorp LP. 2011) using methods described subsequently.

Follow-up period

Outcome data were collected up to 31 March 2014. All participants were assessed for outcomes within 12 months of the date of being accepted to an early intervention service (2026 person-years). Participants with sufficient follow-up data were also assessed for outcomes within 24 months (n=1738; 3476 person-years), 36 months (n=1461; 4383 person-years), 48 months (n=1185; 4740 person-years) and 60 months (n=926; 4630 person-years). Analyses were performed over discrete periods of follow-up rather than using survival analysis owing to non-proportionality of hazards over time for the clinical outcomes described above and in order to facilitate the exploratory mediation analysis described subsequently.

Statistical analysis

Descriptive statistics

Descriptive statistics for predictor, covariate, mediating and outcome variables were obtained as means and SDs for continuous variables (age and number of inpatient days), means and variances for count variables (number of hospital admissions and number of unique antipsychotic medications), and as frequencies and percentages for all other variables.

Associations of cannabis use with demographic factors and clinical outcome

The Mann-Whitney test was used to analyse differences in mean age at presentation (depending on cannabis use) in addition to analysis of age as a categorical variable in regression analyses. Owing to overdispersion (see supplementary material: eTables 1–5), associations with number of hospital admissions and number of unique antipsychotic medications were analysed using multivariable negative binomial regression. Although there was an excess of zero values for number of hospital admissions at 1-year, 2-year and 3-year follow-up, fitting a zero-inflated negative binomial regression model resulted in no meaningful difference compared with standard negative binomial regression (Vuong $p>0.05$ for all models). The association of cannabis use with compulsory hospital admission was assessed using multivariable binary logistic regression. Association with number of inpatient days was assessed using multiple linear regression. Reference groups for covariates in regression analyses were defined as those with the greatest prevalence within each variable. Where covariate data were not recorded (83 participants with unrecorded marital status), this was included as a predictor variable in regression analyses. No patients were dropped from analyses due to missing covariate data.

Mediation of outcomes by antipsychotic treatment failure

In order to test the potential mediation of the effect of cannabis use on outcome variables by antipsychotic treatment failure, an exploratory mediation analysis was performed using the PARAMED module in STATA.[25] This is an extension of the Baron and Kenny method[26] in which a regression model examining the association between the proposed mediator variable and predictor variable is compared with a regression model examining the association between the outcome and the predictor together with the proposed mediator variable. A counterfactual framework which allows for interactions between the exposure and mediator variables is then used to compare the two models to estimate the direct effect of the predictor variable on outcome and the indirect effect of the predictor variable on outcome via the proposed mediator variable.[27] Comparison of the magnitude of the direct and indirect effect allows for estimation of the proportion of total effect that is mediated. In this study, the number of unique antipsychotics (a proxy measure of treatment failure) was selected as a potentially mediating variable of the effect of cannabis use on outcomes (analysed as a linear variable), with age, gender, diagnosis, ethnicity and marital status as covariates. The results are reported as the natural direct effect of cannabis use on outcomes, the natural indirect effect of cannabis use on outcomes mediated by number of unique antipsychotics, and the estimated total effect representing the combined natural direct and indirect effect (figure 1). The percentage of the total effect mediated by number of unique antipsychotics was estimated for the number of days spent in hospital by dividing the natural indirect effect estimate by the total effect and for the number of admissions to hospital and compulsory hospital admission by dividing the natural logarithm of the natural indirect effect by the natural logarithm of the total effect.

RESULTS

Cannabis use among individuals with FEP

Of the total sample, 939 individuals (46.3%) with FEP were found to have a documented history of cannabis use at presentation to early intervention services. Table 1 shows the breakdown of cannabis use by age, gender, ethnicity, marital status and diagnosis. In a multivariable logistic regression analysis (table 1), cannabis use was independently associated with the 16–25-year age group, male gender, single marital status and with a diagnosis of drug-induced psychosis. Cannabis users presented at a younger age than those without documented cannabis use (23.8 vs 24.9 years, Mann-Whitney z=3.84, p<0.001). There was no significant association of cannabis use with ethnicity and cannabis use was less likely among those with psychotic depression or other psychotic disorder not otherwise specified than those with a diagnosis of schizophrenia.

Hospital admission

Figure 2A, B illustrate the mean number of hospital admissions and likelihood of compulsory hospital admission (under the UK MHA) up to 5 years following presentation. Corroborated by multivariable regression analyses (table 2), a recorded history of cannabis use was associated with a significant increase in the number of hospital admissions each year after presentation up to year 5, and a significantly increased likelihood of compulsory hospital admission. The data also showed a greater mean number of days spent in hospital, significant from year 2 onwards, following presentation with a history of cannabis use (figure 2C).

Exploratory mediation analysis

Cannabis use was associated with an increased cumulative likelihood of clozapine (see supplementary material: eTable 4) and number of unique antipsychotics (see supplementary material: eTable 5) prescribed up to 5 years following first presentation. The number of unique antipsychotics prescribed during this period ranged from 0 to 11. While there were no statistically significant differences in clozapine prescription on multivariable logistic regression analysis (see supplementary material: eTable 6), multivariable negative binomial regression (see supplementary material: eTable 7) indicated that a history of cannabis use was associated with an increase in number of unique antipsychotic prescriptions per patient. The exploratory mediation analysis revealed that at 5-year follow-up (table 3) the total effect of cannabis on outcomes was partially mediated by the number of unique antipsychotics prescribed. This was indicated by a significant natural indirect effect of the mediated pathway (figure 1) for each of the three outcomes. The effect of mediation was greatest for the number of days

Figure 1 Mediation analysis.

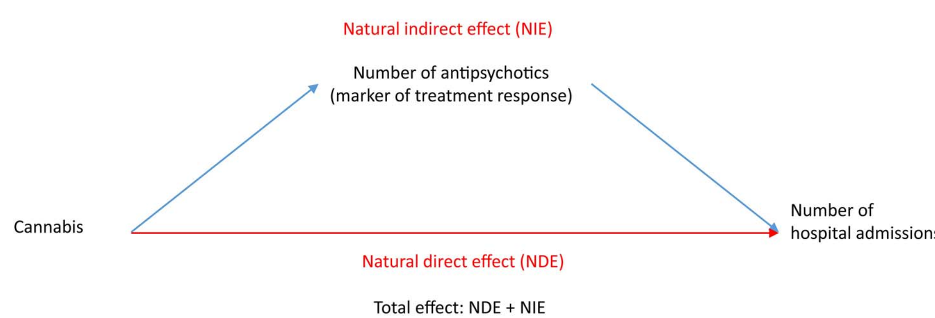


Table 1 Multivariable logistic regression analysis of clinical and demographic factors and history of cannabis use at presentation with first episode psychosis (n=2026)

Factor	Number in sample	Percentage with history of cannabis use (%)	Univariate analysis OR (95% CI), p value	*Multivariable analysis OR (95% CI), p value
Age <16 years	19	10.5	0.11 (0.03 to 0.48), p=0.003	0.12 (0.03 to 0.53), p=0.005
Age 16–25 years	1234	51.7	Reference	Reference
Age 26–35 years	747	39.0	0.60 (0.50 to 0.72), p<0.001	0.70 (0.57 to 0.85), p=0.017
Age >35 years	26	30.8	0.42 (0.18 to 0.96), p=0.04	0.48 (0.20 to 1.14), p=0.006
Female	731	30.5	0.35 (0.29 to 0.43), p<0.001	0.39 (0.32 to 0.48), p<0.001
Male	1295	55.3	Reference	Reference
White	616	49.8	1.21 (0.99 to 1.48), p=0.06	1.17 (0.95 to 1.45), p=0.15
Asian	126	38.9	0.78 (0.53 to 1.13), p=0.19	0.84 (0.56 to 1.25), p=0.38
Black	1005	45.1	Reference	Reference
Other	279	46.6	1.06 (0.81 to 1.39), p=0.65	1.13 (0.84 to 1.50), p=0.42
Married/cohabiting	153	28.8	0.41 (0.28 to 0.59), p<0.001	0.56 (0.38 to 0.82), p=0.003
Divorced/separated	63	23.9	0.32 (0.18 to 0.57), p<0.001	0.47 (0.26 to 0.87), p=0.02
Single	1727	49.4	Reference	Reference
Marital status not recorded	83	31.3	0.47 (0.29 to 0.75), p=0.002	0.50 (0.30 to 0.82), p=0.006
Schizophrenia and related	1097	48.4	Reference	Reference
Bipolar disorder	100	52.0	1.15 (0.77 to 1.74, p=0.49)	1.44 (0.93 to 2.22), p=0.10
Psychotic depression	94	30.9	0.48 (0.30 to 0.75, p=0.001)	0.56 (0.35 to 0.90), p=0.02
Schizoaffective disorder	35	34.2	0.56 (0.27 to 1.13, p=0.10)	0.72 (0.35 to 1.51), p=0.39
Drug-induced psychosis	63	79.0	4.10 (0.62 to 0.92, p<0.001)	3.12 (1.64 to 5.88), p<0.001
Other psychotic disorder	637	41.6	0.76 (0.62 to 0.92, p=0.006)	0.79 (0.64 to 0.97), p=0.02

*Multivariable analysis adjusted for all factors presented in table (and no others).

spent in hospital where number of unique antipsychotics (17.9 days, 95% CI 2.4 to 33.4) mediated 51.4% of the total effect (34.8 days, 95% CI 11.6 to 58.1). Outcomes ascertained at follow-up prior to 5 years (see supplementary material: eTable 8) also indicated similar findings with respect to the mediation effect of number of unique antipsychotics on hospital admission outcomes. However, care should be taken in interpreting these findings owing to the possibility of unmeasured confounding and temporal ambiguity between the mediator and outcome variables.

DISCUSSION

We investigated the impact of cannabis use on outcome as indexed by the number of hospital admissions following onset of illness in a large sample of patients with FEP. The analysis captured data for all 2026 residents who received treatment from early intervention services in four London boroughs over an 8-year timeframe and who had been followed up for up to 5 years.

The use of data recorded in electronic health records presented some challenges in conducting the present study. In particular, the ascertainment of cannabis use was dependent on documentation by a healthcare professional in the course of delivering mental healthcare. Despite this, it was possible to identify cannabis use from electronic health records with a high level of precision. The prevalence of a documented history of cannabis use within 1 month of acceptance by early intervention services was 46.3% in the present study. This is consistent with the high levels of lifetime cannabis use reported in other FEP studies (West London 63%[28]; Cambridge 80.3%[29]). However, it is possible that the prevalence

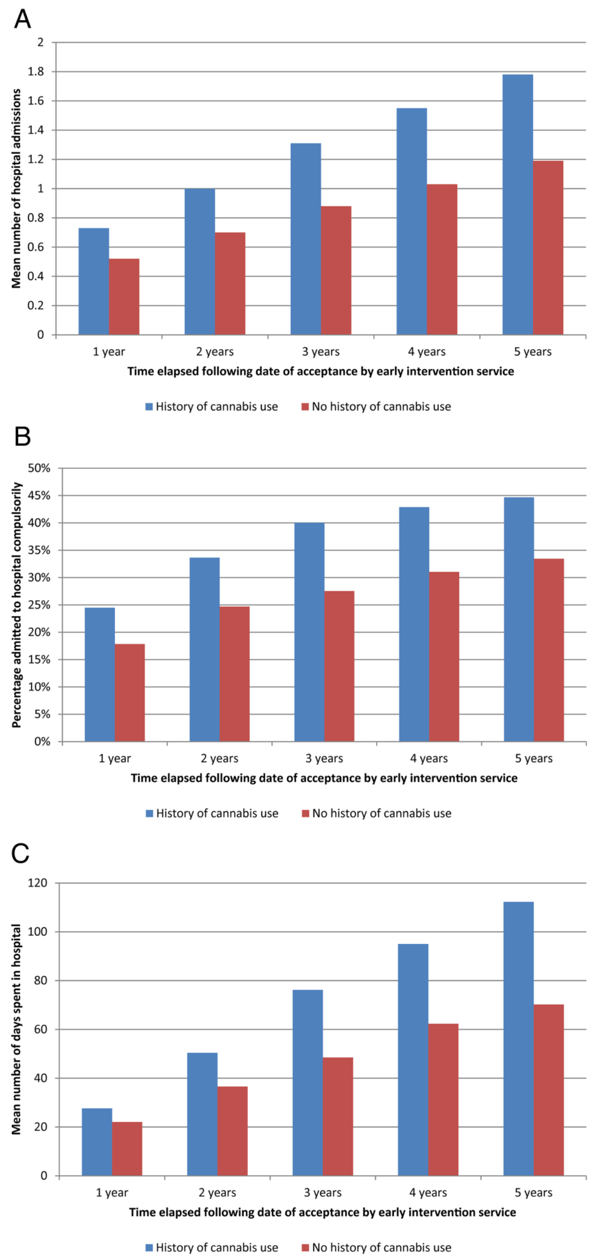


Figure 2 (A) Mean number of hospital admissions among individuals with first episode psychosis with and without documented cannabis use at presentation. (B) Cumulative percentage of patients with first episode psychosis admitted to hospital compulsorily under the UK Mental Act with and without documented cannabis use at presentation. (C) Mean number of days spent in hospital following first episode psychosis depending on history of cannabis use at presentation.

identified in our study underestimated cannabis use owing to under-reporting by patients during clinical assessment. The sample demographic was characteristically young and male, converging with demographic characteristics of other FEP cohorts.[4 10 30 31]

Our findings suggest that patients with a history of cannabis use recorded at presentation to an early intervention service were more likely to be admitted to hospital, to require compulsory admission to hospital, and to spend longer in hospital in the 5 years following presentation. We demonstrated an association between cannabis use and the number of different antipsychotics prescribed during the follow-up period (a proxy marker for treatment failure). Finally, the association between cannabis use and the number of unique antipsychotics was found to mediate the increased risk of subsequent hospitalisation, particularly with respect to number of days spent in hospital.

In the present study, it was not possible to establish on the basis of data recorded in electronic health records whether patients were deemed by clinicians to be resistant to a given antipsychotic following a treatment trial at an adequate dose for an adequate duration before they were changed to another. It is also possible that change to a new antipsychotic may have been prompted by admission to hospital due to a relapse. Nevertheless, change to a different antipsychotic, whether as a result of treatment resistance or poor tolerability, suggests a clinical judgement of failure of treatment with the previous antipsychotic. Regardless of whether the change to a new antipsychotic medication occurred in the community or after an admission to hospital, it is likely that any change in antipsychotic represented a failure of treatment, which must have preceded relapse of illness and hospital admission. Together, these results based on clinical decisions documented by clinicians unbiased by awareness of the objectives of the present study, suggest that cannabis use may be associated with increased risk of hospitalisation in psychosis due to an association with antipsychotic treatment failure. There are a number of ways in which cannabis use may have been associated with antipsychotic treatment failure as suggested by the use of multiple different antipsychotics, including a poor response to treatment, poor adherence to treatment and the presence of adverse side effects. Recent studies have linked a poor response to antipsychotic treatment to the presence of a non-dopaminergic pathophysiology in a subgroup of patients with psychotic disorders.[32] It is possible that increased cannabis use among people with greater number of unique antipsychotics could reflect reduced dopamine synthesis capacity[33] which could reduce response to dopamine receptor blocking antipsychotic medications. Another possibility is that poor medication adherence among such individuals could have an influence on increased number of unique antipsychotics.[5 12] It is noteworthy that in our study, cannabis was associated with an increased likelihood of compulsory hospital admission. A previous study suggests that poor medication adherence is associated with compulsory admission and might also explain its association with cannabis use.[34] While we were not able to tease apart the precise contribution of these various factors to antipsychotic treatment failure, future studies would need to focus on this area, as this may help develop newer strategies for addressing the harmful effects of cannabis.

There are some limitations which should be considered in interpreting the results of this study. The

Table 2 Multivariable analyses of relationship between history of cannabis use at presentation with first episode psychosis and frequency of hospital admissions, likelihood of compulsory hospital admission and mean number of days spent in hospital

Follow-up period	Number in sample	*Number of admissions to hospital Incidence rate ratio (95% CI), p value	†Compulsory hospital admission OR (95% CI), p value	‡Number of days spent in hospital β coefficient (95% CI), p value
1 year	2026	1.37 (1.21 to 1.56), p<0.001	1.33 (1.06 to 1.67), p=0.02	4.1 (−0.6 to 8.7), p=0.09
2 years	1738	1.40 (1.23 to 1.59), p<0.001	1.45 (1.16 to 1.81), p=0.001	9.6 (0.7 to 18.5), p=0.03
3 years	1461	1.48 (1.28 to 1.70), p<0.001	1.65 (1.30 to 2.09), p<0.001	21.6 (8.5 to 34.8), p=0.001
4 years	1185	1.51 (1.29 to 1.76), p<0.001	1.56 (1.20 to 2.02), p=0.001	24.1 (6.1 to 42.0), p=0.009
5 years	926	1.50 (1.25 to 1.80), p<0.001	1.55 (1.16 to 2.08), p=0.003	35.1 (12.1 to 58.1), p=0.003

Results adjusted for age, gender, ethnicity, marital status and psychotic diagnosis.
*Multivariable negative binomial regression.
†Multivariable logistic regression.
‡Multiple linear regression.

findings presented in this study are based on observational, prospectively recorded clinical data. For this reason, it is not possible to infer any aetiological association between cannabis use and greater risk of hospitalisation or treatment failure. However, it would not be feasible or ethical to conduct a randomised controlled trial to investigate the impact of cannabis use on clinical outcomes. We sought to investigate the potential mediation of relapse (indexed by hospital admission) by treatment failure (indexed by the number of unique antipsychotics prescribed). It is possible that switch to a new antipsychotic may have occurred after hospital admission thereby resulting in reversal of mediator and outcome. However, even in cases where switch to a new antipsychotic may have occurred after hospitalisation, it is extremely unlikely that hospital admission triggered the treatment failure that resulted in a need to change antipsychotic therapy. This implies that even in cases where documentation of a change in antipsychotic occurs after hospital admission, the failure of treatment still occurred prior to admission, and so this may have affected the validity of the exploratory mediation analysis resulting in an underestimate of the effect of number of unique antipsychotics on hospital admission outcomes.

Although we sought to adjust multivariable analyses for potentially confounding factors including age, gender, ethnicity, marital status and diagnosis, there may be other unmeasured confounding genetic and environmental factors (including use of alcohol or other illicit substances) which may have influenced the association of cannabis use with outcomes, as well as differences in positive and negative symptom dimensions which we were unable to measure in our study. Unmeasured confounding may also have affected the results from the exploratory mediation analysis investigating number of unique antipsychotics and the association of cannabis on clinical outcomes. For this reason, it is not possible to conclude that antipsychotic treatment failure is the greatest determinant of poor clinical outcomes in relation to cannabis use and there are likely to be other genetic and environmental factors that could influence the effect of cannabis on clinical outcomes in FEP.

In the present study, we investigated the association of cannabis use documented at presentation with FEP with future clinical outcomes. This was defined as cannabis use in clinical documents recorded within 1 month of presentation to early intervention services. Within the first month of presentation, all participants are likely to have undergone a comprehensive clinical assessment allowing systematic ascertainment of documented cannabis use across the whole cohort at inception. However, this method may have underestimated cannabis use owing to under-reporting by patients during clinical assessment. A further bias may have been introduced by selective documentation of assessing clinicians such that documentation of cannabis use was more likely if it was deemed to be of relevance to a patient's clinical presentation.

It is possible that cannabis use varied during the period of follow-up with some people ceasing to use cannabis and others starting to use it. Previous studies suggest that discontinuation of cannabis is associated with improved clinical outcomes in people with FEP[35 36] and bipolar disorder.[37] However, owing to varying level of engagement with mental health services, varying degrees of illness severity and emigration outside the catchment area of clinical services, it was not possible to systematically ascertain ongoing cannabis use in clinical records analysed in this study. It may be that future long-term outcomes were influenced by changes in cannabis use over time. However, if this were the case, it is likely that such variation would have diluted associations with clinical outcomes based on assignment of cannabis use at first presentation to clinical services. It is therefore noteworthy that differences in outcomes based on a history of cannabis use at presentation persisted even at 5-year follow-up. In fact, preliminary analysis of ongoing work in patients with FEP from the same catchment area (n=95) that includes systematic documentation of continuing cannabis use over the follow-up period (by combining clinical records as in the present study with face-to-face research interviews) suggest that 70% of patients with a history of cannabis use at presentation with FEP continued to use cannabis after 3 years, with no new cannabis users who started using following onset of FEP.[38] Hence, taking into consideration the effect of continuing cannabis use would

Table 3 Mediation analysis investigating association of history of cannabis use at presentation with clinical outcomes mediated by number of unique antipsychotics prescribed at 5-year follow-up (n=926)

Outcomes at 5-year follow-up	Natural direct effect (95% CI, p value)	Natural indirect effect (95% CI, p value)	Total effect (95% CI, p value)	Percentage of total effect mediated by number of unique antipsychotics
Number of admissions to hospital: incidence rate ratio	1.37 (1.12 to 1.68, p=0.002)	1.09 (1.01 to 1.18, p=0.03)	1.50 (1.21 to 1.87, p<0.001)	21.3
Compulsory hospital admission: OR	1.39 (0.68 to 2.83, p=0.37)	1.27 (1.03 to 1.58, p=0.03)	1.76 (0.81 to 3.84, p=0.15)	42.3
Number of days spent in hospital: β coefficient (days)	17.0 (−1.5 to 35.4, p=0.07)	17.9 (2.4 to 33.4, p=0.02)	34.8 (11.6 to 58.1, p=0.003)	51.4

Natural direct effect: cannabis use → outcome.
Natural indirect effect: cannabis use → number of unique antipsychotics → outcome.
Total effect: combined natural direct and natural indirect effect.

not have changed the direction of the results reported here, but rather would have demonstrated a stronger adverse effect of cannabis use on outcome in FEP.

Using the cannabis NLP application, it was possible to determine history of cannabis use at presentation with FEP, but it was not possible to determine frequency or amount of cannabis use as this was not systematically recorded in the electronic health record data analysed in this study. Despite this, our findings demonstrated that any cannabis use was significantly associated with poor clinical outcomes, and while the strength of this association may have been greater with increased amount and frequency of cannabis use, such variation is unlikely to have substantially altered the overall association of any cannabis use with poor clinical outcomes that we report here.

These limitations are balanced with the strengths of investigating cannabis use in a large sample of all individuals receiving mental healthcare in early intervention services. Our findings are therefore directly relevant to people who receive care for psychotic disorders in standard clinical practice. The findings presented in this study highlight a clear association between cannabis use and hospitalisation in people with FEP. The fact that over 5 years, cannabis use is associated with 35 additional days spent in hospital has important implications for affected individuals as well healthcare service providers, particularly as almost half of the participants in our study had a history of cannabis use at presentation to early intervention services. This also is the first published study to demonstrate the potential mediation of cannabis use with poorer outcomes by a failure of antipsychotic treatment, albeit with the limitations described previously. Taken together, these findings highlight the importance of ascertaining cannabis use in people receiving care for psychotic disorders and prompt further study to investigate the mechanisms underlying poor clinical outcomes in people who use cannabis and strategies to reduce associated harms.

Author affiliations
[1]Department of Psychosis Studies, King's College London, Institute of Psychiatry, Psychology & Neuroscience, London, UK
[2]Department of Psychological Medicine, King's College London, Institute of Psychiatry, Psychology & Neuroscience, London, UK
[3]South London and Maudsley NHS Foundation Trust, Biomedical Research Centre Nucleus, London, UK

Contributors The study was conceived by SB. The data extraction was led by RP with support from RJ, MBa and HS, supervised by MBr and RS. Statistical analyses were carried out by RP and reporting of findings by RP and RW, supervised by PM and SB. All authors contributed to manuscript preparation and approved the final version.

Funding RJ, MBa, HS, MBr and RS are funded by the National Institute for Health Research (NIHR) Biomedical Research Centre and Dementia Biomedical Research Unit at South London and Maudsley NHS Foundation Trust and King's College London, which also supports the development and maintenance of the BRC Case Register. RP is supported by a UK Medical Research Council (MRC) Clinical Research Training Fellowship (MR/K002813/1). SB has received support from the NIHR (NIHR Clinician Scientist Award; NIHR CS-11-001) and the UK MRC (MR/J012149/1) and from the NIHR Mental Health Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RW is

employed by the Institute of Psychiatry, Psychology & Neuroscience, King's College London.

Disclaimer The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Competing interests All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/coi_disclosure.pdf and declare: The CRIS team (RJ, HS, MBa, MBr and RS) have received research funding from Roche, Pfizer, Johnson & Johnson and Lundbeck. PM has received research funding from Janssen, Sunovion, GW and Roche. SB has received research funding from GW pharmaceuticals.

Ethics approval The CRIS data resource received ethical approval as an anonymised data set for secondary analyses from Oxfordshire REC C (Ref: 08/H0606/71+5).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. United Nations Office on Drugs and Crime. World Drug Report. 2013. http://www.unodc.org/unodc/secured/wdr/wdr2013/World_Drug_Report_2013.pdf
2. Kuepper R, van Os J, Lieb R, *et al*. Continued cannabis use and risk of incidence and persistence of psychotic symptoms: 10 year follow-up cohort study. *BMJ* 2011;342:d738.
3. Moore TH, Zammit S, Lingford-Hughes A, *et al*. Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review. *Lancet* 2007;370:319–28.
4. Hinton M, Edwards J, Elkins K, *et al*. Reductions in cannabis and other illicit substance use between treatment entry and early recovery in patients with first-episode psychosis. *Early Interv Psychiatry* 2007;1:259–66.
5. Zammit S, Moore TH, Lingford-Hughes A, *et al*. Effects of cannabis use on outcomes of psychotic disorders: systematic review. *Br J Psychiatry* 2008;193:357–63.
6. Barrowclough C, Gregg L, Lobban F, *et al*. The impact of cannabis use on clinical outcomes in recent onset psychosis. *Schizophr Bull* 2015;41:382–90.
7. Faber G, Smid HG, Van Gool AR, *et al*. Continued cannabis use and outcome in first-episode psychosis: data from a randomized, open-label, controlled trial. *J Clin Psychiatry* 2012;73:632–8.
8. van Dijk D, Koeter MW, Hijman R, *et al*. Effect of cannabis use on the course of schizophrenia in male patients: a prospective cohort study. *Schizophr Res* 2012;137:50–7.
9. Stone JM, Fisher HL, Major B, *et al*. Cannabis use and first-episode psychosis: relationship with manic and psychotic symptoms, and with age at presentation. *Psychol Med* 2014;44:499–506.
10. Foti DJ, Kotov R, Guey LT, *et al*. Cannabis use and the course of schizophrenia: 10-year follow-up after first hospitalization. *Am J Psychiatry* 2010;167:987–93.
11. Knapp M, Andrew A, McDaid D, *et al*. Investing in recovery: making the business case for effective interventions for people with schizophrenia and psychosis. *Lond Rethink Ment Illn* 2014.
12. Hides L, Dawe S, Kavanagh DJ, *et al*. Psychotic symptom and cannabis relapse in recent-onset psychosis. *Br J Psychiatry* 2006;189:137–43.
13. Lazary J. Psychopharmacological boundaries of schizophrenia with comorbid cannabis use disorder: a critical review. *Curr Pharm Des* 2012;18:4890–6.
14. Barnes TR, Schizophrenia Consensus Group of British Association for Psychopharmacology. Evidence-based guidelines for the pharmacological treatment of schizophrenia: recommendations from the British Association for Psychopharmacology. *J Psychopharmacol* 2011;25:567–620.
15. Stewart R, Soremekun M, Perera G, *et al*. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* 2009;9:51. http://www.biomedcentral.com/1471-244X/9/51 http://dx.doi.org/10.1186/1471-244X-9-51
16. Gafoor R, Nitsch D, McCrone P, *et al*. Effect of early intervention on 5-year outcome in non-affective psychosis. *Br J Psychiatry* 2010;196:372–6.
17. Chang CK, Hayes R, Broadbent M, *et al*. All-cause mortality among people with serious mental illness (SMI), substance use disorders, and depressive disorders in southeast London: a cohort study. *BMC Psychiatry* 2010;10:77.
18. Chang CK, Hayes RD, Perera G, *et al*. Life expectancy at birth for people with serious mental illness and other major disorders from a secondary mental health care case register in London. *PLoS ONE* 2011;6:e19590.
19. Wu CY, Chang CK, Hayes RD, *et al*. Clinical risk assessment rating and all-cause mortality in secondary mental healthcare: the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) Case Register. *Psychol Med* 2012;42:1581–90.
20. Hayes RD, Chang CK, Fernandes AC, *et al*. Functional status and all-cause mortality in serious mental illness. *PLoS ONE* 2012;7:e44613.
21. Jackson R. *TextHunter*. Published Online First: 2 August 2014. http://dx.doi.org/10.5281/zenodo.11122
22. Jackson RG, Ball M, Patel R, *et al*. TextHunter—a user friendly tool for extracting generic concepts from free text in clinical research. *AMIA Annu Symp Proc* 2014;2014:729–38.
23. Mental Health Act. *Great Britain*. London: The Stationery Office, 2007. http://www.legislation.gov.uk/ukpga/2007/12/contents
24. Office for National Statistics. Ethnic group. London: http://www.ons.gov.uk/ons/guide-method/measuring-equality/equality/ethnic-nat-identity-religion/ethnic-group/index.html
25. Emsley R, Liu H. PARAMED: Stata module to perform causal mediation analysis using parametric regression models. *Stat Softw Components* 2013.
26. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 1986;51:1173.
27. Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods* 2013;18:137.
28. Harrison I, Joyce EM, Mutsatsa SH, *et al*. Naturalistic follow-up of co-morbid substance use in schizophrenia: the West London first-episode study. *Psychol Med* 2008;38:79–88.
29. Barnett JH, Werners U, Secher SM, *et al*. Substance use in a population-based clinic sample of people with first-episode psychosis. *Br J Psychiatry* 2007;190:515–20.
30. Carr JA, Norman RM, Manchanda R. Substance misuse over the first 18 months of specialized intervention for first episode psychosis. *Early Interv Psychiatry* 2009;3:221–5.
31. Lambert M, Conus P, Lubman DI, *et al*. The impact of substance use disorders on clinical outcome in 643 patients with first-episode psychosis. *Acta Psychiatr Scand* 2005;112:141–8.
32. Howes OD, Kapur S. A neurobiological hypothesis for the classification of schizophrenia: type A (hyperdopaminergic) and type B (normodopaminergic). *Br J Psychiatry* 2014;205:1–3.
33. Bloomfield MA, Morgan CJ, Egerton A, *et al*. Dopaminergic function in cannabis users and its relationship to cannabis-induced psychotic symptoms. *Biol Psychiatry* 2014;75:470–8.
34. Barbeito S, Vega P, Ruiz de Azúa S, *et al*. Cannabis use and involuntary admission may mediate long-term adherence in first-episode psychosis patients: a prospective longitudinal study. *BMC Psychiatry* 2013;13:326.
35. González-Pinto A, Alberich S, Barbeito S, *et al*. Cannabis and first-episode psychosis: different long-term outcomes depending on continued or discontinued use. *Schizophr Bull* 2011;37:631–9. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=19915168 http://dx.doi.org/10.1093/schbul/sbp126
36. Schoeler T, Monk A, Sami MB, *et al*. Continued versus discontinued cannabis use in patients with psychosis: a systematic review and meta-analysis. *The Lancet Psychiatry* 2016; published online Jan 24. doi:10.1016/S2215-0366(15)00363-6
37. Zorrilla I, Aguado J, Haro JM, *et al*. Cannabis and bipolar disorder: does quitting cannabis use during manic/mixed episode improve clinical/functional outcomes? *Acta Psychiatr Scand* 2015;131:100–10.
38. Schoeler T, Petros N, Behlke I, *et al*. Cannabis use as a predictor for relapse in first episode psychosis: a three-year follow-up study. *Early Interv Psychiatry* 2014;8:58.

Title

Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis: an observational study

Authors

Rashmi Patel BM BCh[1], Robin Wilson MB BS[1], Richard Jackson MSc[2], Michael Ball MSc[2], Hitesh Shetty MSc[3], Matthew Broadbent BSc[3], Robert Stewart MD[2], Philip McGuire MD PhD[1] and Sagnik Bhattacharyya MD PhD[1]

Online supplementary material

eTable 1: Mean number of hospital admissions among individuals with first episode psychosis with and without documented cannabis use at presentation

Follow-up period	History of cannabis use Mean number of admissions (variance, n)	No history of cannabis use Mean number of admissions (variance, n)
1 year, n=2026	0.73 (0.80, n=939)	0.52 (0.70, n=1087)
2 years, n=1738	1.00 (1.36, n=779)	0.70 (1.13, n=959)
3 years, n=1461	1.31 (2.60, n=637)	0.88 (1.74, n=824)
4 years, n=1185	1.55 (3.75, n=492)	1.03 (2.36, n=693)
5 years, n=926	1.78 (4.64, n=367)	1.19 (3.49, n=559)

eTable 2: Cumulative number of patients with first episode psychosis admitted to hospital compulsorily under the UK Mental Health Act with and without documented cannabis use at presentation

Follow-up period	History of cannabis use Compulsory admission n (%)	No history of cannabis use Compulsory admission n (%)
1 year, n=2026	230 (24.5%)	194 (17.9%)
2 years, n=1738	262 (33.6%)	237 (24.7%)
3 years, n=1461	255 (40.0%)	227 (27.6%)
4 years, n=1185	211 (42.9%)	215 (31.0%)
5 years, n=926	164 (44.7%)	187 (33.5%)

eTable 3: Mean number of days spent in hospital following first episode psychosis depending on history of cannabis use

Follow-up period	History of cannabis use Mean number of inpatient days (std dev, n)	No history of cannabis use Mean number of inpatient days (std dev, n)
1 year, n=2026	27.6 (52.2, n=939)	22.0 (50.3, n=1087)
2 years, n=1738	50.4 (98.0, n=779)	36.6 (83.6, n=959)
3 years, n=1461	76.2 (139.7, n=637)	48.5 (107.1, n=824)
4 years, n=1185	95.0 (167.8, n=492)	62.3 (138.6, n=693)
5 years, n=926	112.3 (200.0, n=367)	70.2 (150.3, n=559)

eTable 4: Cumulative number of patients with first episode psychosis with and without documented cannabis use at presentation who were subsequently prescribed clozapine

Follow-up period	History of cannabis use Clozapine n (%)	No history of cannabis use Clozapine n (%)
1 year, n=2026	29 (3.1%)	33 (3.0%)
2 years, n=1738	54 (6.9%)	57 (5.9%)
3 years, n=1461	69 (10.8%)	64 (7.8%)
4 years, n=1185	63 (12.8%)	76 (11.0%)
5 years, n=926	58 (15.8%)	67 (12.0%)

eTable 5: Mean number of uniquely prescribed antipsychotics among individuals with first episode psychosis with and without documented cannabis use at presentation

Follow-up period	History of cannabis use Mean number of antipsychotics (variance, n)	No history of cannabis use Mean number of antipsychotics (variance, n)
1 year, n=2026	1.67 (1.37, n=939)	1.49 (1.34, n=1087)
2 years, n=1738	2.05 (2.14, n=779)	1.80 (2.09, n=959)
3 years, n=1461	2.30 (2.77, n=637)	1.98 (2.70, n=824)
4 years, n=1185	2.47 (3.28, n=492)	2.20 (3.22, n=693)
5 years, n=926	2.68 (3.98, n=367)	2.34 (3.66, n=559)

eTable 6: Multivariable logistic regression analysis of association between history of cannabis use at presentation with first episode psychosis and clozapine prescription during follow-up period

Follow-up period	Clozapine prescription Odds ratio (95% CI, p value)
1 year, n=2026	0.90 (0.53 to 1.53, p=0.69)
2 years, n=1738	1.00 (0.66 to 1.49, p=0.98)
3 years, n=1461	1.32 (0.90 to 1.92, p=0.15)
4 years, n=1185	0.99 (0.68 to 1.44, p=0.95)
5 years, n=926	1.14 (0.77 to 1.71, p=0.51)
Results adjusted for age, gender, ethnicity, marital status and psychotic diagnosis	

eTable 7: Multivariable negative binomial regression analysis of association between history of cannabis use at presentation with first episode psychosis and number of unique antipsychotic medications prescribed during follow-up period

Follow-up period	Number of unique antipsychotics prescribed Incidence rate ratio (95% CI, p value)
1 year, n=2026	1.13 (1.05 to 1.21), p=0.001
2 years, n=1738	1.11 (1.03 to 1.19), p=0.004
3 years, n=1461	1.15 (1.06 to 1.24), p=0.001
4 years, n=1185	1.09 (1.00 to 1.19), p=0.05
5 years, n=926	1.13 (1.02 to 1.25), p=0.02
Results adjusted for age, gender, ethnicity, marital status and psychotic diagnosis	

eTable 8: Mediation analysis investigating association of history of cannabis use at presentation with clinical outcomes mediated by number of unique antipsychotics prescribed during follow-up period

Outcome variable	Follow-up period	Natural direct effect (95% CI, p value)	Natural indirect effect (95% CI, p value)	Total effect (95% CI, p value)
Number of admissions to hospital Incidence rate ratio	1 year, n=2026	1.33 (1.11 to 1.60, p=0.002)	1.06 (1.03 to 1.09, p=0.001)	1.41 (1.17 to 1.70, p<0.001)
	2 years, n=1738	1.35 (1.14 to 1.60, p<0.001)	1.06 (1.02 to 1.11, p=0.006)	1.44 (1.21 to 1.71, p<0.001)
	3 years, n=1461	1.35 (1.16 to 1.58, p<0.001)	1.10 (1.04 to 1.16, p=0.001)	1.49 (1.26 to 1.76, p<0.001)
	4 years, n=1185	1.42 (1.22 to 1.66, p<0.001)	1.07 (1.00 to 1.14, p-0.07)	1.52 (1.28 to 1.80, p<0.001)
	5 years, n=926	1.37 (1.12 to 1.68, p=0.002)	1.09 (1.01 to 1.18, p=0.03)	1.50 (1.21 to 1.87, p<0.001)
Compulsory hospital admission Odds ratio	1 year, n=2026	1.10 (0.75 to 1.62, p=0.63)	1.15 (1.06 to 1.24, p=0.001)	1.26 (0.85 to 1.88, p=0.26)
	2 years, n=1738	1.13 (0.71 to 1.81, p=0.60)	1.15 (1.04 to 1.27, p=0.006)	1.30 (0.80 to 2.12, p=0.29)
	3 years, n=1461	1.42 (0.94 to 2.15, p=0.10)	1.25 (1.09 to 1.43, p=0.002)	1.77 (1.12 to 2.80, p=0.02)
	4 years, n=1185	1.57 (0.89 to 2.77, p=0.12)	1.18 (0.99 to 1.41, p=0.07)	1.85 (1.00 to 3.41, p=0.05)
	5 years, n=926	1.39 (0.68 to 2.83, p=0.37)	1.27 (1.03 to 1.58, p=0.03)	1.76 (0.81 to 3.84, p=0.15)
Number of days spent in hospital B coefficient (days)	1 year, n=2026	0.3 (-3.9 to 4.4, p=0.90)	3.8 (1.7 to 5.9, p<0.001)	4.1 (-0.6 to 8.7, p=0.09)
	2 years, n=1738	2.2 (-5.4 to 9.7, p=0.57)	7.4 (2.2 to 12.5, p=0.005)	9.6 (0.6 to 18.5, p=0.04)
	3 years, n=1461	8.0 (-2.9 to 18.9, p=0.15)	13.6 (5.3 to 21.8, p<0.001)	21.6 (8.4 to 34.8, p<0.001)
	4 years, n=1185	13.2 (-1.3 to 27.7, p=0.07)	10.8 (-0.7 to 22.2, p=0.07)	24.0 (5.9 to 42.1, p=0.009)
	5 years, n=926	17.0 (-1.5 to 35.4, p=0.07)	17.9 (2.4 to 33.4, p=0.02)	34.8 (11.6 to 58.1, p=0.003)

Natural direct effect: cannabis use → outcome
Natural indirect effect: cannabis use → number of unique antipsychotics → outcome
Total effect: combined natural direct and natural indirect effect

4.2 Supplementary methods

4.21 SQL data extraction

In this study, a cohort of patients presenting to Early Intervention Services for First Episode Psychosis was selected. The SQL script below illustrates the method for selecting this cohort from the BRC Case Register.

```
USE [SQLCRIS_User]
select  a.brcid,
                   accepted_date,
                   case
                            when accepted_date between '1 Apr 2006' and '31 Mar
2007' then '2006-7'
                            when accepted_date between '1 Apr 2007' and '31 Mar
2008' then '2007-8'
                            when accepted_date between '1 Apr 2008' and '31 Mar
2009' then '2008-9'
                            when accepted_date between '1 Apr 2009' and '31 Mar
2010' then '2009-10'
                            when accepted_date between '1 Apr 2010' and '31 Mar
2011' then '2010-11'
                            when accepted_date between '1 Apr 2011' and '31 Mar
2012' then '2011-12'
                            when accepted_date between '1 Apr 2012' and '31 Mar
2013' then '2012-13'
                   end as 'accepted_year',
                   case
                            when accepted_date between '1 Apr 2006' and '31 Mar
2007' then '1'
                            when accepted_date between '1 Apr 2007' and '31 Mar
2008' then '2'
                            when accepted_date between '1 Apr 2008' and '31 Mar
2009' then '3'
                            when accepted_date between '1 Apr 2009' and '31 Mar
2010' then '4'
                            when accepted_date between '1 Apr 2010' and '31 Mar
2011' then '5'
                            when accepted_date between '1 Apr 2011' and '31 Mar
2012' then '6'
                            when accepted_date between '1 Apr 2012' and '31 Mar
2013' then '7'
                   end as 'accepted_year_recode',
                   location_name,
```

```
                case
                        when
                                Location_Name like '%COAST%' or
                                Location_Name like '%Leo%' or
                                Location_Name like '%Lewisham Early%' or
                                Location_Name like 'STEP'
                        then '1' else '0'
                end as 'EIS',
                case
                        when
                                Location_Name like '%OASIS%'
                        then '1' else '0'
                end as 'OASIS',
                floor((datediff (day, p.cleaneddateofbirth , cast
(accepted_date as datetime))/365)) age,
                p.Gender_ID  Gender,
                p.ethnicitycleaned ethnicity,
                p.Marital_Status_ID marital_status,
                p.Employment_ID Employment_status,
                p.Housing_Status Accommodation_status
from (
        select distinct brcid,accepted_date,Location_Name,ROW_NUMBER()
        over (partition by brcid order by accepted_date) spell
        from [brhnsql094].[SQLCRIS].[dbo].[Team_episode] r
        where
        Accepted_Date between '1 Apr 2006' and '31 Mar 2013'
        and Referral_Admin_Status_ID not like 'rejected'
        and (
                Location_Name like '%COAST%' or
                Location_Name like '%Leo%' or
                Location_Name like '%Lewisham Early%' or
                Location_Name like '%OASIS%' or
                Location_Name like 'STEP')
    ) a
left join sqlcris.dbo.epr_form p on a.brcid=p.brcid
where spell=1
```

In this SQL query, the first accepted referral to an Early Intervention Service or high risk clinical service was selected by performing a sub-query in the FROM statement. The base table for selecting patients was “[SQLCRIS].[dbo].[Team_episode]”. As described in section 2.5, this is a table with data on accepted and rejected referrals to SLaM clinical teams. The currency of this table is the “team episode” which represents one referral to a particular SLaM team. Each patient may therefore have many team episodes. For this reason, it is necessary to select only one team episode in order to generate a query whose currency is individual patients (i.e. unique BRCIDs). To do this, the sub-query selects patients whose referrals were accepted ("Referral_Admin_Status_ID not like 'rejected'") between 1st April 2006 and 31st March 2014 ("Accepted_Date between '1 Apr 2006' and '31 Mar 2013'") to an Early Intervention Service or high risk clinical service (COAST, LEO, Lewisham Early, OASIS or STEP). The referrals were partitioned by each patient ("partition by brcid") and ordered in ascending date of acceptance ("order by accepted_date") to generate a "spell" column containing an integer number greater than or equal to "1". Only the first team episode fulfilling these criteria were selected ("where spell=1").

From this cohort of patients (all BRCIDs in table "a"), the date of being accepted to the team was selected ("accepted_date") and recoded using CASE WHEN statements into years from 2006/7 to 2012/13. Further CASE WHEN statements were employed to determine whether patients were accepted to an Early Intervention Service ("EIS") or a high risk clinical service ("OASIS"). A table join between "a" and the EPR form as table "p" was performed ("left join sqlcris.dbo.epr_form p on a.brcid=p.brcid") in order to extract demographic data for each patient (age, gender, ethnicity, marital status, employment status and accommodation status).

The output of this query was saved as a view in the following location: "[SQLCRIS_User].[RPatel].[rp_EIScohort]" from which the next script was applied.

```
USE [SQLCRIS_User]
SELECT r.[brcid]
      ,r.[accepted_date]
      ,r.[accepted_year]
      ,r.[accepted_year_recode]
      ,r.[location_name]
      ,b.[diagnosis_date]
      ,b.[primary_diagnosisrecode]
        ,case
              when (
                      b.[primary_diagnosisrecode] like 'F2xSchizophrenia')
              then '1'
```

```
                when (
                        b.[primary_diagnosisrecode] like 'Bipolar')
                then '2'

                when (
                        b.[primary_diagnosisrecode] like
'PsychoticDepression')
                then '3'

                when (
                        b.[primary_diagnosisrecode] like 'Schizoaffective')
                then '4'

                when (
                        b.[primary_diagnosisrecode] like 'F1x.5DrugPsyc')
                then '5'

                when (
                        b.[primary_diagnosisrecode] like 'OtherPsychosis')
                then '6'

                end as Firstpsychosisdiagnosisrecode2
      ,r.[EIS]
      ,r.[OASIS]
      ,r.[age]
      ,r.[Gender]
      ,case
                when
                        r.Gender like 'Female'
                then '0'

                when
                        r.Gender like 'Male'
                then '1'

                when
                        r.Gender like 'Not Known'
                then '2'

                end as GenderRecode2
      ,r.[ethnicity]
      ,case
                when (
                        r.ethnicity like 'Irish (B)' or
                        r.ethnicity like 'Any other white background (C)' or
                        r.ethnicity like 'British (A)')
                then 'White'
```

```
            when (
                    r.ethnicity like 'Bangladeshi (K)' or
                    r.ethnicity like 'Pakistani (J)' or
                    r.ethnicity like 'Chinese (R)' or
                    r.ethnicity like 'Indian (H)' or
                    r.ethnicity like 'Any other Asian background (L)')
            then 'Asian'

            when (
                    r.ethnicity like 'African (N)')
            then 'BlackAfrican'

            when (
                    r.ethnicity like 'Caribbean (M)')
            then 'BlackCaribbean'

            when (
                    r.ethnicity like 'Any other black background (P)' or
                    r.ethnicity like 'White and Black African (E)')
            then 'BlackOther'

            when (
                    r.ethnicity like 'White and Asian (F)' or
                    r.ethnicity like 'Any other mixed background (G)' or
                    r.ethnicity like 'Any other ethnic group (S)' or
                    r.ethnicity like 'Not Stated (Z)' or
                    r.ethnicity like 'None')
            then 'Other'

            end as ethnicrecode,

    case
            when (
                    r.ethnicity like 'Irish (B)' or
                    r.ethnicity like 'Any other white background (C)' or
                    r.ethnicity like 'British (A)')
            then '1'

            when (
                    r.ethnicity like 'Bangladeshi (K)' or
                    r.ethnicity like 'Pakistani (J)' or
                    r.ethnicity like 'Chinese (R)' or
                    r.ethnicity like 'Indian (H)' or
                    r.ethnicity like 'Any other Asian background (L)')
            then '2'

            when (
                    r.ethnicity like 'African (N)')
            then '3'
```

```
                when (
                        r.ethnicity like 'Caribbean (M)')
                then '4'

                when (
                        r.ethnicity like 'Any other black background (P)' or
                        r.ethnicity like 'White and Black African (E)')
                then '5'

                when (
                        r.ethnicity like 'White and Asian (F)' or
                        r.ethnicity like 'Any other mixed background (G)' or
                        r.ethnicity like 'Any other ethnic group (S)' or
                        r.ethnicity like 'Not Stated (Z)' or
                        r.ethnicity like 'None')
                then '6'

                end as ethnicrecode2
      ,r.[marital_status]
      ,case
                when (
                        r.marital_status like 'Cohabiting' or
                        r.marital_status like 'Married' or
                        r.marital_status like 'Married/Civil Partner')
                then 'MarriedCohabiting'

                when (
                        r.marital_status like 'Divorced' or
                        r.marital_status like 'Divorced/Civil Partnership
Dissolved' or
                        r.marital_status like 'Separated')
                then 'DivorcedSeparated'

                when (
                        r.marital_status like 'Single')
                then 'Single'

                when (
                        r.marital_status like 'Widowed' or
                        r.marital_status like 'Widowed/Surviving Civil
Partner')
                then 'Widowed'

                when (
                        r.marital_status like 'Not Disclosed' or
                        r.marital_status like 'Not Known')
                then 'NotRecorded'
```

```
                end as maritalrecode,

      case
                when (
                        r.marital_status like 'Cohabiting' or
                        r.marital_status like 'Married' or
                        r.marital_status like 'Married/Civil Partner')
                then '1'

                when (
                        r.marital_status like 'Divorced' or
                        r.marital_status like 'Divorced/Civil Partnership
Dissolved' or
                        r.marital_status like 'Separated')
                then '2'

                when (
                        r.marital_status like 'Single')
                then '3'

                when (
                        r.marital_status like 'Widowed' or
                        r.marital_status like 'Widowed/Surviving Civil
Partner')
                then '4'

                when (
                        r.marital_status like 'Not Disclosed' or
                        r.marital_status like 'Not Known')
                then '5'

                end as maritalrecode2
      ,r.[Employment_status]
      ,case
                when (
                        r.Employment_status like 'Volunteer' or
                        r.Employment_status like 'Self Employed' or
                        r.Employment_status like 'Part Time Employment' or
                        r.Employment_status like 'Paid Employment')
                then 'Employed'

                when (
                        r.Employment_status like 'Govt Training Scheme' or
                        r.Employment_status like 'Full Time Student' or
                        r.Employment_status like 'Full Time Student - School
age')
                then 'Student'

                when (
```

```
                r.Employment_status like 'Retired')
        then 'Retired'

        when (
                r.Employment_status like 'Registered Disabled' or
                r.Employment_status like 'Unemployed')
        then 'Unemployed'

        when (
                r.Employment_status like 'Other' or
                r.Employment_status like 'Not Known' or
                r.Employment_status like 'xNx')
        then 'NotRecorded'

        end as employmentrecode,

    case
        when (
                r.Employment_status like 'Volunteer' or
                r.Employment_status like 'Self Employed' or
                r.Employment_status like 'Part Time Employment' or
                r.Employment_status like 'Paid Employment')
        then '1'

        when (
                r.Employment_status like 'Govt Training Scheme' or
                r.Employment_status like 'Full Time Student' or
                r.Employment_status like 'Full Time Student - School
age')
        then '2'

        when (
                r.Employment_status like 'Retired')
        then '3'

        when (
                r.Employment_status like 'Registered Disabled' or
                r.Employment_status like 'Unemployed')
        then '4'

        when (
                r.Employment_status like 'Other' or
                r.Employment_status like 'Not Known' or
                r.Employment_status like 'xNx')
        then '5'

        end as employmentrecode2
    ,r.[Accommodation_status]
    ,case
```

```
                when (
                        r.Accommodation_status like 'Owner')
                then 'Owner'

                when (
                        r.Accommodation_status like 'Private Tenant')
                then 'PrivateTenant'

                when (
                        r.Accommodation_status like 'Council Tenant')
                then 'CouncilTenant'

                when (
                        r.Accommodation_status like 'Trust' or
                        r.Accommodation_status like 'Nursing/Residential')
                then 'SupportedAccommodation'

                when (
                        r.Accommodation_status like 'Homeless')
                then 'Homeless'

                when (
                        r.Accommodation_status like 'Other')
                then 'Other'

                when (
                        r.Accommodation_status like 'Not known' or
                        r.Accommodation_status like 'xNx')
                then 'NotRecorded'

                end as accommodationrecode,

        case
                when (
                        r.Accommodation_status like 'Owner')
                then '1'

                when (
                        r.Accommodation_status like 'Private Tenant')
                then '2'

                when (
                        r.Accommodation_status like 'Council Tenant')
                then '3'

                when (
                        r.Accommodation_status like 'Trust' or
                        r.Accommodation_status like 'Nursing/Residential')
                then '4'
```

```
        when (
                r.Accommodation_status like 'Homeless')
        then '5'

        when (
                r.Accommodation_status like 'Other')
        then '6'

        when (
                r.Accommodation_status like 'Not known' or
                r.Accommodation_status like 'xNx')
        then '7'

        end as accommodationrecode2
,case when (
        select distinct brcid
        from sqlcris.dbo.mha_section
        where
                Start_Date between r.accepted_date and
                DATEADD(d,14,r.accepted_date) and
                brcid=r.brcid
        ) IS  null then 0 else 1 end  mhasection2w,
case when (
        select distinct brcid
        from sqlcris.dbo.mha_section
        where
                Start_Date between r.accepted_date and
                DATEADD(m,1,r.accepted_date) and
                brcid=r.brcid
        ) IS  null then 0 else 1 end  mhasection1m,
case when (
        select distinct brcid
        from sqlcris.dbo.mha_section
        where
                Start_Date between r.accepted_date and
                DATEADD(m,3,r.accepted_date) and
                brcid=r.brcid
        ) IS  null then 0 else 1 end  mhasection3m,
case when (
        select distinct brcid
        from sqlcris.dbo.mha_section
        where
                Start_Date between r.accepted_date and
                DATEADD(m,6,r.accepted_date) and
                brcid=r.brcid
        ) IS  null then 0 else 1 end  mhasection6m,
case when (
        select distinct brcid
```

```
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.accepted_date and
                        DATEADD(m,12,r.accepted_date) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end  mhasection12m,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.accepted_date and
                        DATEADD(m,24,r.accepted_date) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end  mhasection24m,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.accepted_date and
                        DATEADD(m,36,r.accepted_date) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end  mhasection36m,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.accepted_date and
                        DATEADD(m,48,r.accepted_date) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end  mhasection48m,
        case when (
                select distinct brcid
                from sqlcris.dbo.mha_section
                where
                        Start_Date between r.accepted_date and
                        DATEADD(m,60,r.accepted_date) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end  mhasection60m,
        case when (
                select distinct brcid
                from sqlcris.dbo.inpatient_episode
                where
                        admission_Date between r.accepted_date and
                        DATEADD(d,14,r.accepted_date) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end  inpatient2w,
        case when (
                select distinct brcid
                from sqlcris.dbo.inpatient_episode
```

```
            where
                    admission_Date between r.accepted_date and
                    DATEADD(m,1,r.accepted_date) and
                    brcid=r.brcid
            ) IS  null then 0 else 1 end  inpatient1m,
    case when (
            select distinct brcid
            from sqlcris.dbo.inpatient_episode
            where
                    admission_Date between r.accepted_date and
                    DATEADD(m,3,r.accepted_date) and
                    brcid=r.brcid
            ) IS  null then 0 else 1 end  inpatient3m,
    case when (
            select distinct brcid
            from sqlcris.dbo.inpatient_episode
            where
                    admission_Date between r.accepted_date and
                    DATEADD(m,6,r.accepted_date) and
                    brcid=r.brcid
            ) IS  null then 0 else 1 end  inpatient6m,
    case when (
            select distinct brcid
            from sqlcris.dbo.inpatient_episode
            where
                    admission_Date between r.accepted_date and
                    DATEADD(m,12,r.accepted_date) and
                    brcid=r.brcid
            ) IS  null then 0 else 1 end  inpatient12m,
    case when (
            select distinct brcid
            from sqlcris.dbo.inpatient_episode
            where
                    admission_Date between r.accepted_date and
                    DATEADD(m,24,r.accepted_date) and
                    brcid=r.brcid
            ) IS  null then 0 else 1 end  inpatient24m,
    case when (
            select distinct brcid
            from sqlcris.dbo.inpatient_episode
            where
                    admission_Date between r.accepted_date and
                    DATEADD(m,36,r.accepted_date) and
                    brcid=r.brcid
            ) IS  null then 0 else 1 end  inpatient36m,
    case when (
            select distinct brcid
            from sqlcris.dbo.inpatient_episode
            where
```

```
                        admission_Date between r.accepted_date and
                        DATEADD(m,48,r.accepted_date) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end  inpatient48m,
        case when (
                select distinct brcid
                from sqlcris.dbo.inpatient_episode
                where
                        admission_Date between r.accepted_date and
                        DATEADD(m,60,r.accepted_date) and
                        brcid=r.brcid
                ) IS  null then 0 else 1 end  inpatient60m,
                (SELECT         COUNT(*) AS noofadmissions
                        FROM    sqlcris.dbo.inpatient_episode
                        WHERE   (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,12,r.accepted_date)) AND
                                        (Discharge_Date >= r.accepted_date)
OR
                                        (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,12,r.accepted_date)) AND
                                        (Discharge_Date = '1 Jan 1900')
                        GROUP BY BrcId) AS noofadmissions_12m,
                (SELECT         COUNT(*) AS noofadmissions
                        FROM    sqlcris.dbo.inpatient_episode
                        WHERE   (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,24,r.accepted_date)) AND
                                        (Discharge_Date >= r.accepted_date)
OR
                                        (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,24,r.accepted_date)) AND
                                        (Discharge_Date = '1 Jan 1900')
                        GROUP BY BrcId) AS noofadmissions_24m,
                (SELECT         COUNT(*) AS noofadmissions
                        FROM    sqlcris.dbo.inpatient_episode
                        WHERE   (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,36,r.accepted_date)) AND
                                        (Discharge_Date >= r.accepted_date)
OR
                                        (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,36,r.accepted_date)) AND
                                        (Discharge_Date = '1 Jan 1900')
                        GROUP BY BrcId) AS noofadmissions_36m,
                (SELECT         COUNT(*) AS noofadmissions
```

```
                        FROM    sqlcris.dbo.inpatient_episode
                        WHERE   (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,48,r.accepted_date)) AND
                                        (Discharge_Date >= r.accepted_date)
OR
                                        (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,48,r.accepted_date)) AND
                                        (Discharge_Date = '1 Jan 1900')
                        GROUP BY BrcId) AS noofadmissions_48m,
                (SELECT         COUNT(*) AS noofadmissions
                        FROM    sqlcris.dbo.inpatient_episode
                        WHERE   (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,60,r.accepted_date)) AND
                                        (Discharge_Date >= r.accepted_date)
OR
                                        (BrcId = r.brcid) AND
                                        (Admission_Date <=
DATEADD(m,60,r.accepted_date)) AND
                                        (Discharge_Date = '1 Jan 1900')
                        GROUP BY BrcId) AS noofadmissions_60m,
                SQLCrisImport.dbo.getlos(accepted_date, DATEADD(m, 12,
accepted_date), r.brcid) AS los_12m,
                SQLCrisImport.dbo.getlos(accepted_date, DATEADD(m, 24,
accepted_date), r.brcid) AS los_24m,
                SQLCrisImport.dbo.getlos(accepted_date, DATEADD(m, 36,
accepted_date), r.brcid) AS los_36m,
                SQLCrisImport.dbo.getlos(accepted_date, DATEADD(m, 48,
accepted_date), r.brcid) AS los_48m,
                SQLCrisImport.dbo.getlos(accepted_date, DATEADD(m, 60,
accepted_date), r.brcid) AS los_60m,
      case when (
                select distinct brcid
                from
[GateDB_Cris].[RPatel].[vw_gate_hunter_cannabis_filteredmatch] b
                where b.prob >0.7524131 and (
                        b.mlObservation1 like 'positive')
                        and
                        Convert(datetime,b.Document_Date,103) between
DATEADD(m,-1,r.accepted_date) and DATEADD(m,1,r.accepted_date)
                        and
                        b.brcid=r.brcid
                ) IS not null then 1 else 0 end Cannabis1m,
                (select MAX(drugorder)
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
```

```
                        Start_Date <= DATEADD(d,14,r.accepted_date) and
                        brcid=r.brcid
                group by brcid) as antipsychoticcount2w,
                (select MAX(drugorder)
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,1,r.accepted_date) and
                        brcid=r.brcid
                group by brcid) as antipsychoticcount1m,
                (select MAX(drugorder)
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,3,r.accepted_date) and
                        brcid=r.brcid
                group by brcid) as antipsychoticcount3m,
                (select MAX(drugorder)
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,6,r.accepted_date) and
                        brcid=r.brcid
                group by brcid) as antipsychoticcount6m,
                (select MAX(drugorder)
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,12,r.accepted_date) and
                        brcid=r.brcid
                group by brcid) as antipsychoticcount12m,
                (select MAX(drugorder)
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,24,r.accepted_date) and
                        brcid=r.brcid
                group by brcid) as antipsychoticcount24m,
                (select MAX(drugorder)
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,36,r.accepted_date) and
                        brcid=r.brcid
                group by brcid) as antipsychoticcount36m,
                (select MAX(drugorder)
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
```

```
                        Start_Date <= DATEADD(m,48,r.accepted_date) and
                        brcid=r.brcid
                group by brcid) as antipsychoticcount48m,
                (select MAX(drugorder)
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,60,r.accepted_date) and
                        brcid=r.brcid
                group by brcid) as antipsychoticcount60m,
        case when (
                select distinct brcid
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(d,14,r.accepted_date) and
                        drugrecoded like 'Clozapine' and
                        brcid=r.brcid
                ) IS null then 0 else 1 end Clozapine2w,
        case when (
                select distinct brcid
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,1,r.accepted_date) and
                        drugrecoded like 'Clozapine' and
                        brcid=r.brcid
                ) IS null then 0 else 1 end Clozapine1m,
        case when (
                select distinct brcid
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,3,r.accepted_date) and
                        drugrecoded like 'Clozapine' and
                        brcid=r.brcid
                ) IS null then 0 else 1 end Clozapine3m,
        case when (
                select distinct brcid
                from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
                where
                        Start_Date <= DATEADD(m,6,r.accepted_date) and
                        drugrecoded like 'Clozapine' and
                        brcid=r.brcid
                ) IS null then 0 else 1 end Clozapine6m,
        case when (
                select distinct brcid
```

```
            from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
            where
                    Start_Date <= DATEADD(m,12,r.accepted_date) and
                    drugrecoded like 'Clozapine' and
                    brcid=r.brcid
            ) IS null then 0 else 1 end Clozapine12m,
    case when (
            select distinct brcid
            from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
            where
                    Start_Date <= DATEADD(m,24,r.accepted_date) and
                    drugrecoded like 'Clozapine' and
                    brcid=r.brcid
            ) IS null then 0 else 1 end Clozapine24m,
    case when (
            select distinct brcid
            from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
            where
                    Start_Date <= DATEADD(m,36,r.accepted_date) and
                    drugrecoded like 'Clozapine' and
                    brcid=r.brcid
            ) IS null then 0 else 1 end Clozapine36m,
    case when (
            select distinct brcid
            from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
            where
                    Start_Date <= DATEADD(m,48,r.accepted_date) and
                    drugrecoded like 'Clozapine' and
                    brcid=r.brcid
            ) IS null then 0 else 1 end Clozapine48m,
    case when (
            select distinct brcid
            from
[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]
            where
                    Start_Date <= DATEADD(m,60,r.accepted_date) and
                    drugrecoded like 'Clozapine' and
                    brcid=r.brcid
            ) IS null then 0 else 1 end Clozapine60m
  FROM [SQLCRIS_User].[RPatel].[rp_EIScohort] r
left join [SQLCRIS_User].[RPatel].[rp_EIScohort_diagafter_recode_first] b
on r.brcid=b.brcid
where OASIS=0
```

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

In this query, various sources of data were joined to the cohort of patients defined in "[SQLCRIS_User].[RPatel].[rp_EIScohort]". A left join was performed on the "[SQLCRIS_User].[RPatel].[rp_EIScohort_diagafter_recode_first]" view as "b" in order to extract the first recorded psychotic disorder diagnosis ("b.[primary_diagnosisrecode]") and the date it was recorded ("b.[diagnosis_date]"). These were recoded into diagnostic groups by means of a CASE WHEN statement into the "Firstpsychosisdiagnosisrecode2" column. Further CASE WHEN statements were employed to recode demographic data into suitable categories for statistical analysis. CASE WHEN statements with sub-queries were performed to extract outcome data between 2 weeks and 5 years after being accepted to a SLAM clinical service using the same methods described in section 3.21. These included whether patients had been compulsory admitted to hospital under the UK Mental Health Act ("mhasectionXX"), any psychiatric hospital admission ("inpatientXX") and the number of psychiatric hospital admissions ("noofadmissions_XXX") obtained through a "COUNT" sub-query. The number of days spent in hospital during the follow-up period was obtained using a SQL store procedure ("SQLCrisImport.dbo.getlos").

The presence of cannabis use was obtained by means of a CASE WHEN statement containing a sub-query on the "[GateDB_Cris].[RPatel].[vw_gate_hunter_cannabis_filteredmatch]" view described in further detail in section 4.22. In this query, NLP data current or historical exposure to cannabis use were obtained by specifying an SVM marginal threshold to obtain a minimum precision of 90% ("b.prob >0.7524131") including positive instances of cannabis exposure ("b.mlObservation1 like 'positive'") which were documented within one month of being accepted to a SLaM clinical service ("Convert(datetime,b.Document_Date,103) between DATEADD(m,-1,r.accepted_date) and DATEADD(m,1,r.accepted_date)"). The output of this sub-query was stored in the "Cannabis1m" column.

The final section specifies CASE WHEN statements with sub-queries on the "[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]" table described in further detail in section 4.22. The first set of CASE WHEN statements extracted the number of unique antipsychotics ("MAX(drugorder)") occurring within the follow-up period ("antipsychoticcountXX"). The second set of CASE WHEN statements extracted whether the patient was exposed to clozapine during the follow-up period ("ClozapineXX").

A final WHERE statement was added to discard patients accepted to the OASIS service ("where OASIS=0") thereby producing a dataset only including patients presenting to Early Intervention Services for first-episode psychosis.

4.22 SQL support queries

In order to determine the number of unique antipsychotics prescribed during the follow-up period, a SQL query was generated to provide a cumulative count of the number of different antipsychotics each patient in the BRC Case Register had been exposed to. The base SQL query to achieve this is reproduced below.

```
USE [SQLCRIS_User]
SELECT a.[brcid]
      ,[source_table]
      ,[start_date]
      ,[drug]
      ,[drugrecoded]
      ,[cn_doc_id]
  FROM [SQLCRIS_User].[RPatel].[rp_EIScohort] a
  left join
[SQLCRIS_User].[RPatel].[rp_medication_combined_antipsychotic_recode] b on
a.brcid=b.brcid
  where start_date>=a.accepted_date and drugrecoded not like 'OtherDrug'
```

This query performed a left join to the "[SQLCRIS_User].[RPatel].[rp_medication_combined_antipsychotic_recode]" view described in section 2.52. This view recodes medication data in the BRC Case Register to create categories representing different antipsychotic medications based on their generic and trade names. A WHERE statement was specified to select medications started after presenting to a SLaM clinical service ("start_date>=a.accepted_date") and to filter out non-antipsychotic medications ("drugrecoded not like 'OtherDrug'").

The output of this query was saved as a view in the following location: "[SQLCRIS_User].[RPatel].[rp_EIScohort_Antipsychoticafter]" from which the next script was applied.

```
SELECT [brcid]
      ,ROW_NUMBER() OVER(PARTITION BY brcid ORDER BY start_date ASC)
drugorder
      ,[start_date]
      ,[drugrecoded]
FROM (
               SELECT  brcid,
                               start_date,
                               drugrecoded,
                               ROW_NUMBER() OVER(PARTITION BY
drugrecoded,brcid ORDER BY start_date ASC) rn
                       FROM
[SQLCRIS_User].[RPatel].[rp_EIScohort_Antipsychoticafter]
                       ) a
WHERE rn = 1
```

In this SQL query, the ROW_NUMBER command was used to identify the number of antipsychotic medications started over time. This was achieved by performing a sub-query in the FROM statement to first partition by the name of the antipsychotic medication ("drugrecoded") and the patient ID ("brcid") ordering by ascending order of start date ("ORDER BY start_date ASC"). This sub-query groups the individual instances of medication data by the name of the drug and then the patient. By specifying a WHERE statement with "rn = 1", only the first instance of each drug is preserved. In this way, multiple records of the same drug are discarded. This allows the main query, "ROW_NUMBER() OVER(PARTITION BY brcid ORDER BY start_date ASC) drugorder" to then order each unique drug according to their start date. By way of example, a patient started on olanzapine and then switched to risperidone and then aripiprazole may have the following data represented in the

“[SQLCRIS_User].[RPatel].[rp_EIScohort_Antipsychoticafter]” view:

brcid	start_date	drugrecoded	rn	drugorder
4568135	02/05/2006	Olanzapine	1	1
4568135	05/05/2006	Olanzapine	2	1
4568135	14/07/2006	Risperidone	1	2
4568135	15/07/2006	Risperidone	2	2
4568135	19/07/2006	Risperidone	3	2
4568135	24/07/2006	Risperidone	4	2
4568135	15/09/2006	Risperidone	5	2
4568135	18/04/2008	Aripiprazole	1	3
4568135	24/04/2008	Aripiprazole	2	3
4568135	26/04/2008	Aripiprazole	3	3
4568135	30/04/2008	Aripiprazole	4	3
4568135	02/05/2008	Aripiprazole	5	3

Running the query described above would reduce the output to the following:

brcid	start_date	drugrecoded	drugorder
4568135	02/05/2006	Olanzapine	1
4568135	14/07/2006	Risperidone	2
4568135	18/04/2008	Aripiprazole	3

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

The output of this query was saved as a table in the following location:

“[SQLCRIS_User].[RPatel].[tbl_rp_EIScohort_Antipsychoticafter_collapsed]” from which the SQL query described in section 4.21 was applied.

In order to extract data on psychotic disorder diagnosis, a SQL query was performed to obtain all diagnosis recorded after the date of acceptance to a SLaM clinical service which were consistent with the presence of a psychotic disorder.

```
USE [SQLCRIS_User]
SELECT a.[brcid]
      ,[source_table]
      ,[diagnosis_date]
      ,[primary_diagnosis]
      ,[cn_doc_id]
  FROM [SQLCRIS_User].[RPatel].[rp_EIScohort] a
  left join [SQLCrisImport].[dbo].[diagnosis_combined] b on a.brcid=b.brcid
  where diagnosis_date>=a.accepted_date and
(primary_diagnosis LIKE '%psychotic%' OR
primary_diagnosis LIKE '%psychosis%' OR
primary_diagnosis LIKE '%with psyc%' OR
primary_diagnosis LIKE '%schizophreni%' OR
primary_diagnosis LIKE '%scizophreni%' OR
primary_diagnosis LIKE '%schizotyp%' OR
primary_diagnosis LIKE '%scizotyp%' OR
primary_diagnosis LIKE '%delusion%' OR
primary_diagnosis LIKE '%hallucin%' OR
primary_diagnosis LIKE '%thought disorder%' OR
primary_diagnosis LIKE '%first rank%' OR
primary_diagnosis LIKE '%schizoaffect%' OR
primary_diagnosis LIKE '%scizoaffect%' OR
primary_diagnosis LIKE '%mania%' OR
primary_diagnosis LIKE '%manic%' OR
primary_diagnosis LIKE '%f20%' OR
primary_diagnosis LIKE '%f21%' OR
primary_diagnosis LIKE '%f22%' OR
primary_diagnosis LIKE '%f23%' OR
primary_diagnosis LIKE '%f24%' OR
primary_diagnosis LIKE '%f25%' OR
primary_diagnosis LIKE '%f28%' OR
primary_diagnosis LIKE '%f29%' OR
primary_diagnosis LIKE '%f30%' OR
primary_diagnosis LIKE '%f31.2%' OR
```

```
primary_diagnosis LIKE '%f32.3%' OR
primary_diagnosis LIKE '%f33.3%') AND
primary_diagnosis NOT LIKE '%without psyc%' AND
primary_diagnosis NOT LIKE '%personality%' AND
primary_diagnosis NOT LIKE '%trichotil%' AND
primary_diagnosis NOT LIKE '%kleptomani%' AND
primary_diagnosis NOT LIKE '%hypomani%'
```

In this SQL query, a left join was performed between “[SQLCRIS_User].[RPatel].[rp_EIScohort]” and “[SQLCrisImport].[dbo].[diagnosis_combined]” with a WHERE statement to include diagnoses occurring after the date of acceptance to a SLaM clinical service (“diagnosis_date>=a.accepted_date”) and indicating a psychotic disorder diagnosis.

The output of this query was saved as a view in the following location: “[SQLCRIS_User].[RPatel].[rp_EIScohort_diagafter]” from which the subsequent SQL query was applied.

```
USE [SQLCRIS_User]
select
        [brcid],
        [source_table],
    [diagnosis_date],
    [primary_diagnosis] primary_diagnosisraw,
    [cn_doc_id],
        case
                when (
                        primary_diagnosis like '%schizophreni%' or
                        primary_diagnosis like '%scizophreni%' or
                        primary_diagnosis like '%schizotyp%' or
                        primary_diagnosis like '%scizotyp%' or
                        primary_diagnosis like '%f20%' or
                        primary_diagnosis like '%f21%' or
                        primary_diagnosis like '%f22%' or
                        primary_diagnosis like '%f23%' or
                        primary_diagnosis like '%f24%' or
                        primary_diagnosis like '%f28%' or
                        primary_diagnosis like '%f29%')
                then 'F2xSchizophrenia'
```

```
        when (
                primary_diagnosis like '%schizoaffect%' or
                primary_diagnosis like '%scizoaffect%' or
                primary_diagnosis like '%f25%')
        then 'Schizoaffective'

        when (
                primary_diagnosis like '%f30%' or
                primary_diagnosis like '%f31%' or
                primary_diagnosis like '%manic%' or
                primary_diagnosis like '%mania%' or
                primary_diagnosis like '%bipolar%' or
                primary_diagnosis like '%bpad%' or
                primary_diagnosis like '%affective disorder%' or
                primary_diagnosis like '%mixed affective%') and
                primary_diagnosis not like '%trichotillomania%' and
                primary_diagnosis not like '%kleptomania%'
        then 'Bipolar'

        when (
                primary_diagnosis like '%psychosis%' or
                primary_diagnosis like '%psychotic%' or
                primary_diagnosis like '%with psyc%' or
                primary_diagnosis like '%f32.3%' or
                primary_diagnosis like '%f33.3%') and
                (primary_diagnosis like '%depress%' or
                primary_diagnosis like '%f32%' or
                primary_diagnosis like '%f33%') and
                primary_diagnosis not like '%without%'
        then 'PsychoticDepression'

        when (
                primary_diagnosis like '%drug%' or
                primary_diagnosis like '%alcohol%' or
                primary_diagnosis like '%opioid%' or
                primary_diagnosis like '%opiate%' or
                primary_diagnosis like '%cannabi%' or
                primary_diagnosis like '%benzo%' or
                primary_diagnosis like '%hallucinogen%' or
                primary_diagnosis like '%cocaine%' or
                primary_diagnosis like '%cannabis%' or
                primary_diagnosis like '%f10%' or
                primary_diagnosis like '%f11%' or
                primary_diagnosis like '%f12%' or
                primary_diagnosis like '%f13%' or
                primary_diagnosis like '%f14%' or
                primary_diagnosis like '%f15%' or
                primary_diagnosis like '%f16%' or
                primary_diagnosis like '%f17%' or
```

```
                       primary_diagnosis like '%f18%' or
                       primary_diagnosis like '%f19%')
               then 'F1x.5DrugPsyc'
               else 'OtherPsychosis'
               end as primary_diagnosisrecode
  FROM [SQLCRIS_User].[RPatel].[rp_EIScohort_diagafter]
```

This query employed a series of CASE WHEN statements to recode the psychotic disorder diagnoses using a method analogous to that described in section 3.21.

The output of this query was saved as a view in the following location: "[SQLCRIS_User].[RPatel].[rp_EIScohort_diagafter_recode]" from which the subsequent SQL query was applied.

```
USE [SQLCRIS_User]
SELECT brcid,
      diagnosis_date,
      primary_diagnosisrecode
FROM (
               SELECT  brcid,
                              diagnosis_date,
                              primary_diagnosisrecode,
                              ROW_NUMBER() OVER(PARTITION BY brcid ORDER
BY diagnosis_date ASC) rn
                       FROM RPatel.rp_EIScohort_diagafter_recode
                       ) a

WHERE rn = 1
order by brcid, diagnosis_date
```

This query selected the first recorded psychotic disorder diagnosis using a sub-query in the FROM statement which employed the ROW_NUMBER command to order the diagnosis records by ascending date and a WHERE statement to select the first recorded diagnosis ("rn = 1").

The output of this query was saved as a view in the following location:

“[SQLCRIS_User].[RPatel].[rp_EIScohort_diagafter_recode_first]” from which the SQL query described in section 4.21 was applied.

4.23 Cannabis NLP development

In order to ascertain cannabis use, an NLP application was developed to extract documentation of current or historical cannabis use within one month of presenting to an Early Intervention Service. A dictionary of key words was specified to extract all sentences containing a potential reference to cannabis use. These were as follows: cannab; marij; marih; weed; pot; ganja; grass; resin; hash; skunk; THC; CBD; spice; K2. These key words were chosen to cover medical and slang terms related to herbal cannabis as well as synthetic cannabinoids (“spice” and “K2”). All key words were selected as wildcards in order to account for misspellings and compound words (e.g. cannabis, cannabidiol, tetrahydrocannabinol). As a result, some sentences contained key words which were irrelevant to cannabis use. In order to filter out these sentences, a SQL query was written to manually remove sentences containing irrelevant key words as follows:

```
USE [GateDB_Cris]
GO
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE view [RPatel].[vw_gate_hunter_cannabis_filteredmatch] as
SELECT *
  FROM [GateDB_Cris].[RPatel].[gate_hunter_cannabis_new]
  where
  match not like 'Hospice' and
  match not like 'Spicer' and
  match not like 'Gillgrass' and
  match not like 'Hashim' and
  match not like 'HEALTHCARE' and
  match not like 'weeding' and
  match not like 'Marije' and
  match not like 'Hashmi' and
  match not like 'auspices' and
```

```
  match not like 'Gilgrass' and
  match not like 'Rozemarijn' and
  match not like 'FORTHCOMING' and
  match not like 'Tweed' and
  match not like 'Nathasha' and
  match not like 'incresing' and
  match not like 'grassing' and
  match not like 'Marija' and
  match not like 'incresingly' and
  match not like 'Hashemi' and
  match not like 'expresing' and
  match not like 'Tweedy' and
  match not like 'Ingrassia' and
  match not like 'hashimoto' and
  match not like 'Hashieka' and
  match not like 'resinded' and
  match not like 'Takahashi-' and
  match not like 'weeded' and
  match not like 'spiced' and
  match not like 'weedkiller' and
  match not like 'Gringrass' and
  match not like 'Shasha' and
  match not like 'grassy' and
  match not like 'Shashi' and
  match not like 'distresing' and
  match not like 'dresing' and
  match not like 'Marijke' and
  match not like 'Hashad' and
  match not like 'Sweeden' and
  match not like 'Hashimi' and
  match not like 'Weedon' and
  match not like 'hospices' and
  match not like 'Hashima' and
  match not like 'trinityhospice' and
  match not like 'Grassmere' and
  match not like 'spicey' and
  match not like 'resing' and
  match not like 'resind' and
  match not like 'Depresin' and
  match not like 'Rehashed' and
  match not like 'Shash' and
  match not like 'gjithashtu' and
  match not like 'Hashimotos' and
  match not like 'seaweed' and
  match not like 'presing' and
  match not like 'Revieweed' and
  match not like 'progresing' and
  match not like 'Saruhashi' and
  match not like 'weed-killer' and
```

```
match not like 'grasshopper' and
match not like 'Subhash' and
match not like 'addresing' and
match not like 'weedy' and
match not like 'Hashi' and
match not like 'aggrassive' and
match not like 'Takahashi' and
match not like 'Marijia' and
match not like 'rehashing' and
match not like 'jDspicer' and
match not like 'Showeed' and
match not like 'rehash' and
match not like 'decresing' and
match not like 'Weednesday' and
match not like 'Hashtroudi' and
match not like 'Hasheem' and
match not like 'Hashi-' and
match not like 'Manjubhashini' and
match not like 'Gharghasht' and
match not like 'Snodgrass' and
match not like 'Hashir' and
match not like 'Hashes' and
match not like 'alloweed' and
match not like 'Desmopresin' and
match not like 'Grassarah' and
match not like 'Capgrass' and
match not like 'Grassby' and
match not like 'Hashtrodi' and
match not like 'Grassa' and
match not like 'Ambresin' and
match not like 'cannabalism' and
match not like 'reviweed' and
match not like 'dhashay' and
match not like 'sweedish' and
match not like 'resinstated' and
match not like 'auspice' and
match not like 'Kinshasha' and
match not like 'Marijka' and
match not like 'Lanchashire' and
match not like 'spicemen' and
match not like 'Amarijit' and
match not like 'Weeder' and
match not like 'followeed' and
match not like 'Dishashi' and
match not like 'weedend' and
match not like 'Hashma' and
match not like 'grasshoppers' and
match not like 'resinding' and
match not like 'Hashims' and
```

```
match not like 'diuresing' and
match not like 'Gingrass' and
match not like 'grassland' and
match not like 'Shashee' and
match not like 'Tweedie' and
match not like 'Lemongrass' and
match not like 'Tweedlie' and
match not like 'Kamishasherene' and
match not like 'EALTHCARE' and
match not like 'Hasher' and
match not like 'Grasso' and
match not like 'Hashmita' and
match not like 'resinstate' and
match not like 'HEATHCOTE' and
match not like 'Tweedle' and
match not like 'Elaweed' and
match not like 'Grassroots' and
match not like 'resined' and
match not like 'thashe' and
match not like 'hashed' and
match not like 'answeed' and
match not like 'Desmopresine' and
match not like 'csnodgrass' and
match not like 'Resinente' and
match not like 'Tweedale' and
match not like 'Marijah' and
match not like 'resindment' and
match not like 'Hairdresing' and
match not like 'Hashard' and
match not like 'Anne-marije'
```

After filtering out irrelevant key words, all sentences within event and correspondence notes in the BRC Case Register containing any one (or more) of these key words was extracted using TextHunter software. A random selection of 233 sentences was chosen and annotated independently by two human annotators (Rashmi Patel and Michael Ball) to form the reference “gold” dataset against which the resulting NLP application was tested. Inter-annotator agreement between the two annotators was tested. There was 88% agreement between the two annotators with a Cohen’s kappa value of 0.76. A further 478 sentences were annotated by Rashmi Patel. These formed the “seed” training dataset for SVM NLP development. The best performing NLP application derived using 10 fold cross validation from the seed dataset was found to have a baseline precision of 81%

and recall of 95% against the gold reference dataset. However, applying an SVM margin filter to yield a minimum precision of 90% resulted in a drop of recall to 45%. For this reason, a further two rounds of active learning performed by Rashmi Patel on an additional 1357 sentences resulted in an improvement of precision statistics with a minimum precision of 90% associated with recall of 72%. This is summarised in Table 2 and Figure 1 (magenta, yellow and grey ROC curves) of Jackson et al[119] reproduced in the chapter 6 of this thesis.

4.24 Statistical analysis

The STATA commands used to perform the analyses described in the main manuscript and supplementary material are reproduced below.

```
**Recoding demographic variables to deal with missing data and regrouping categories**
gen agegp=age
recode agegp min/15=1 16/25=2 26/35=3 36/max=4
gen ethnicrecode3 = ethnicrecode2
recode ethnicrecode3 1=1 2=2 3=3 4=3 5=3 6=4 7=5
gen ethnicrecode3b = ethnicrecode3
recode ethnicrecode3b 5=.
gen maritalrecode3 = maritalrecode2
recode maritalrecode3 1=1 2=2 3=3 4=3 5=4
gen maritalrecode3b = maritalrecode3
recode maritalrecode3b 4=.
gen employmentrecode3 = employmentrecode2
recode employmentrecode3 1=1 2=2 3=3 4=3 5=4
gen employmentrecode3b = employmentrecode3
recode employmentrecode3b 4=.
gen accommodationrecode3 = accommodationrecode2
recode accommodationrecode3 7=6
gen accommodationrecode3b = accommodationrecode2
recode accommodationrecode3b 7=.
gen fullcovariate1 = 1
replace fullcovariate1 = 0 if ethnicrecode3b==.|maritalrecode3b==.
gen fullcovariate2 = 1
replace fullcovariate2 = 0 if
ethnicrecode3b==.|maritalrecode3b==.|employmentrecode3b==.|accommodationrec
ode3b==.

**Label variables and create ordinal exposures**
```

```
label define Firstpsychosisdiagnosisrecode2 1 "F2xSchizophrenia" 2
"Bipolar" 3 "PsychoticDepression" 4 "Schizoaffective" 5 "F1x.5DrugPsyc" 6
"OtherPsychosis"
label define agegp 1 "<16" 2 "16-25" 3 "26-35" 4 ">35"
label define AgeRecode2 1 "<16" 2 "16-24" 3 "24-49" 4 "50-64" 5 ">64"
label define GenderRecode2 0 "Female" 1 "Male" 2 "NotRecorded"
label define ethnicrecode2 1 "White" 2 "Asian" 3 "BlackAfrican" 4
"BlackCaribbean" 5 "BlackOther" 6 "Other" 7 "NotRecorded"
label define ethnicrecode3 1 "White" 2 "Asian" 3 "Black" 4 "Other" 5
"NotRecorded"
label define ethnicrecode3b 1 "White" 2 "Asian" 3 "Black" 4 "Other"
label define maritalrecode2 1 "MarriedCohabiting" 2 "DivorcedSeparated" 3
"Single" 4 "Widowed" 5 "NotRecorded"
label define maritalrecode3 1 "MarriedCohabiting" 2 "DivorcedSeparated" 3
"Single" 4 "NotRecorded"
label define maritalrecode3b 1 "MarriedCohabiting" 2 "DivorcedSeparated" 3
"Single"
label define employmentrecode2 1 "Employed" 2 "Student" 3 "Retired" 4
"Unemployed" 5 "NotRecorded"
label define employmentrecode3 1 "Employed" 2 "Student" 3 "Unemployed" 4
"NotRecorded"
label define employmentrecode3b 1 "Employed" 2 "Student" 3 "Unemployed"
label define accommodationrecode2 1 "Owner" 2 "PrivateTenant" 3
"CouncilTenant" 4 "SupportedAccomodation" 5 "Homeless" 6 "Other" 7
"NotRecorded"
label define accommodationrecode3 1 "Owner" 2 "PrivateTenant" 3
"CouncilTenant" 4 "SupportedAccomodation" 5 "Homeless" 6 "Other"
label define accommodationrecode3b 1 "Owner" 2 "PrivateTenant" 3
"CouncilTenant" 4 "SupportedAccomodation" 5 "Homeless" 6 "Other"
label define accepted_year_recode 1 "2006-7" 2 "2007-8" 3 "2008-9" 4 "2009-
10" 5 "2010-11" 6 "2011-12" 7 "2012-13"
label values Firstpsychosisdiagnosisrecode2 Firstpsychosisdiagnosisrecode2
label values agegp agegp
label values AgeRecode2 AgeRecode2
label values GenderRecode2 GenderRecode2
label values ethnicrecode2 ethnicrecode2
label values ethnicrecode3 ethnicrecode3
label values ethnicrecode3b ethnicrecode3b
label values maritalrecode2 maritalrecode2
label values maritalrecode3 maritalrecode3
label values maritalrecode3b maritalrecode3b
label values employmentrecode2 employmentrecode2
label values employmentrecode3 employmentrecode3
label values employmentrecode3b employmentrecode3b
label values accommodationrecode2 accommodationrecode2
label values accommodationrecode3 accommodationrecode3
label values accommodationrecode3b accommodationrecode3b
label values accepted_year_recode accepted_year_recode
```

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

```
**Descriptive statistics**

tab fullcovariate1
tab Cannabis1m
summ age, detail
summ age if Cannabis1m==0, detail
summ age if Cannabis1m==1, detail
ksmirnov age, by(Cannabis1m)
ranksum age, by(Cannabis1m)
tab Cannabis1m Firstpsychosisdiagnosisrecode2, row chi2
**tab Cannabis1m Firstpsychosisdiagnosisrecode2, row exact**
tab Cannabis1m Secondpsydiagnosisrecode2, row chi2
**tab Cannabis1m Secondpsydiagnosisrecode2, row exact**
tab Firstpsychosisdiagnosisrecode2 Secondpsydiagnosisrecode2, row chi2
**tab Firstpsychosisdiagnosisrecode2 Secondpsydiagnosisrecode2, row exact**
tab Cannabis1m agegp, row chi2
tab Cannabis1m agegp, row exact
tab Cannabis1m GenderRecode2, row chi2
tab Cannabis1m GenderRecode2, row exact
tab Cannabis1m ethnicrecode3b, row chi2
tab Cannabis1m ethnicrecode3b, row exact
tab Cannabis1m maritalrecode3b, row chi2
tab Cannabis1m maritalrecode3b, row exact

**Demographic variables including missing data**
tab Cannabis1m maritalrecode3, row chi2
tab Cannabis1m maritalrecode3, row exact

**Demographic variables logistic regression**
logistic Cannabis1m i.Firstpsychosisdiagnosisrecode2
logistic Cannabis1m i.Secondpsydiagnosisrecode2
logistic Cannabis1m ib2.agegp
logistic Cannabis1m ib1.GenderRecode2
logistic Cannabis1m ib3.ethnicrecode3
logistic Cannabis1m ib3.maritalrecode3b
logistic Cannabis1m i.Firstpsychosisdiagnosisrecode2 ib2.agegp
ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3b

**Demographic variables logistic regression including missing**
logistic Cannabis1m ib3.maritalrecode3
logistic Cannabis1m i.Firstpsychosisdiagnosisrecode2 ib2.agegp
ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3

**MHA Section**
tab Cannabis1m mhasection12m, row
tab Cannabis1m mhasection24m if accepted_year_recode<=6, row
tab Cannabis1m mhasection36m if accepted_year_recode<=5, row
tab Cannabis1m mhasection48m if accepted_year_recode<=4, row
tab Cannabis1m mhasection60m if accepted_year_recode<=3, row
```

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

```
**Antipsychotic count**
summ antipsychoticcount12m if Cannabis1m==0, detail
summ antipsychoticcount12m if Cannabis1m==1, detail
summ antipsychoticcount24m if Cannabis1m==0 & accepted_year_recode<=6,
detail
summ antipsychoticcount24m if Cannabis1m==1 & accepted_year_recode<=6,
detail
summ antipsychoticcount36m if Cannabis1m==0 & accepted_year_recode<=5,
detail
summ antipsychoticcount36m if Cannabis1m==1 & accepted_year_recode<=5,
detail
summ antipsychoticcount48m if Cannabis1m==0 & accepted_year_recode<=4,
detail
summ antipsychoticcount48m if Cannabis1m==1 & accepted_year_recode<=4,
detail
summ antipsychoticcount60m if Cannabis1m==0 & accepted_year_recode<=3,
detail
summ antipsychoticcount60m if Cannabis1m==1 & accepted_year_recode<=3,
detail

**Clozapine**
tab Cannabis1m Clozapine12m, row
tab Cannabis1m Clozapine24m if accepted_year_recode<=6, row
tab Cannabis1m Clozapine36m if accepted_year_recode<=5, row
tab Cannabis1m Clozapine48m if accepted_year_recode<=4, row
tab Cannabis1m Clozapine60m if accepted_year_recode<=3, row

**Number of admissions**
summ noofadmissions_12m if Cannabis1m==0, detail
summ noofadmissions_12m if Cannabis1m==1, detail
summ noofadmissions_24m if Cannabis1m==0 & accepted_year_recode<=6, detail
summ noofadmissions_24m if Cannabis1m==1 & accepted_year_recode<=6, detail
summ noofadmissions_36m if Cannabis1m==0 & accepted_year_recode<=5, detail
summ noofadmissions_36m if Cannabis1m==1 & accepted_year_recode<=5, detail
summ noofadmissions_48m if Cannabis1m==0 & accepted_year_recode<=4, detail
summ noofadmissions_48m if Cannabis1m==1 & accepted_year_recode<=4, detail
summ noofadmissions_60m if Cannabis1m==0 & accepted_year_recode<=3, detail
summ noofadmissions_60m if Cannabis1m==1 & accepted_year_recode<=3, detail

**Number of inpatient days**
summ los_12m if Cannabis1m==0, detail
summ los_12m if Cannabis1m==1, detail
summ los_24m if Cannabis1m==0 & accepted_year_recode<=6, detail
summ los_24m if Cannabis1m==1 & accepted_year_recode<=6, detail
summ los_36m if Cannabis1m==0 & accepted_year_recode<=5, detail
summ los_36m if Cannabis1m==1 & accepted_year_recode<=5, detail
summ los_48m if Cannabis1m==0 & accepted_year_recode<=4, detail
summ los_48m if Cannabis1m==1 & accepted_year_recode<=4, detail
```

```
summ los_60m if Cannabis1m==0 & accepted_year_recode<=3, detail
summ los_60m if Cannabis1m==1 & accepted_year_recode<=3, detail

**Multivariable analyses with missing data included**

**MHA section**
logistic mhasection12m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3
logistic mhasection24m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=6
logistic mhasection36m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=5
logistic mhasection48m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=4
logistic mhasection60m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=3

**Antipsychotic count - nbreg**
nbreg antipsychoticcount12m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3, irr
nbreg antipsychoticcount24m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=6, irr
nbreg antipsychoticcount36m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=5, irr
nbreg antipsychoticcount48m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=4, irr
nbreg antipsychoticcount60m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=3, irr

**Number of admissions - nbreg**
nbreg noofadmissions_12m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3, irr
nbreg noofadmissions_24m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=6, irr
nbreg noofadmissions_36m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=5, irr
nbreg noofadmissions_48m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=4, irr
```

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

```
nbreg noofadmissions_60m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=3, irr

**Inpatient days**
regress los_12m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2 ib2.agegp
ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3
regress los_24m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2 ib2.agegp
ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=6
regress los_36m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2 ib2.agegp
ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=5
regress los_48m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2 ib2.agegp
ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=4
regress los_60m i.Cannabis1m i.Firstpsychosisdiagnosisrecode2 ib2.agegp
ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=3

**Medition analysis**

**Antipsychotic count predictor**
logistic mhasection12m antipsychoticcount12m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3
logistic mhasection24m antipsychoticcount24m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3
logistic mhasection36m antipsychoticcount36m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3
logistic mhasection48m antipsychoticcount48m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3
logistic mhasection60m antipsychoticcount60m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3

nbreg noofadmissions_12m antipsychoticcount12m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3, irr
nbreg noofadmissions_24m antipsychoticcount24m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=6, irr
nbreg noofadmissions_36m antipsychoticcount36m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=5, irr
```

```
nbreg noofadmissions_48m antipsychoticcount48m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=4, irr
nbreg noofadmissions_60m antipsychoticcount60m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=3, irr

regress los_12m antipsychoticcount12m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3
regress los_24m antipsychoticcount24m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=6
regress los_36m antipsychoticcount36m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=5
regress los_48m antipsychoticcount48m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=4
regress los_60m antipsychoticcount60m i.Firstpsychosisdiagnosisrecode2
ib2.agegp ib1.GenderRecode2 ib3.ethnicrecode3 ib3.maritalrecode3 if
accepted_year_recode<=3

**Adding in antipsychoticcount as covariate**
logistic mhasection12m i.Cannabis1m antipsychoticcount12m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3
logistic mhasection24m i.Cannabis1m antipsychoticcount24m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3
logistic mhasection36m i.Cannabis1m antipsychoticcount36m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3
logistic mhasection48m i.Cannabis1m antipsychoticcount48m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3
logistic mhasection60m i.Cannabis1m antipsychoticcount60m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3

nbreg noofadmissions_12m i.Cannabis1m antipsychoticcount12m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3, irr
nbreg noofadmissions_24m i.Cannabis1m antipsychoticcount24m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=6, irr
nbreg noofadmissions_36m i.Cannabis1m antipsychoticcount36m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=5, irr
```

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

```
nbreg noofadmissions_48m i.Cannabis1m antipsychoticcount48m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=4, irr
nbreg noofadmissions_60m i.Cannabis1m antipsychoticcount60m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=3, irr

regress los_12m i.Cannabis1m antipsychoticcount12m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3
regress los_24m i.Cannabis1m antipsychoticcount24m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=6
regress los_36m i.Cannabis1m antipsychoticcount36m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=5
regress los_48m i.Cannabis1m antipsychoticcount48m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=4
regress los_60m i.Cannabis1m antipsychoticcount60m
i.Firstpsychosisdiagnosisrecode2 ib2.agegp ib1.GenderRecode2
ib3.ethnicrecode3 ib3.maritalrecode3 if accepted_year_recode<=3

**Generate shorter variable names and dummy variables**
gen antipsy2w = antipsychoticcount2w
gen antipsy1m = antipsychoticcount1m
gen antipsy3m = antipsychoticcount3m
gen antipsy6m = antipsychoticcount6m
gen antipsy12m = antipsychoticcount12m
gen antipsy24m = antipsychoticcount24m
gen antipsy36m = antipsychoticcount36m
gen antipsy48m = antipsychoticcount48m
gen antipsy60m = antipsychoticcount60m
xi i.agegp i.GenderRecode2

**Mediation analysis - paramed**
**N.B. paramed cannot use "if"*

paramed noofadmissions_12m, avar(Cannabis1m) mvar(antipsy12m) a0(0) a1(1)
m(0) yreg(negbin) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
paramed mhasection12m, avar(Cannabis1m) mvar(antipsy12m) a0(0) a1(1) m(0)
yreg(logistic) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
```

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

```
paramed los_12m, avar(Cannabis1m) mvar(antipsy12m) a0(0) a1(1) m(0)
yreg(linear) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4 _IGenderRec_1
_IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5 _IFirstpsyc_6
_Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2 _Imaritalre_3
_Imaritalre_4)

drop if accepted_year_recode>=7
paramed noofadmissions_24m, avar(Cannabis1m) mvar(antipsy24m) a0(0) a1(1)
m(0) yreg(negbin) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
paramed mhasection24m, avar(Cannabis1m) mvar(antipsy24m) a0(0) a1(1) m(0)
yreg(logistic) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
paramed los_24m, avar(Cannabis1m) mvar(antipsy24m) a0(0) a1(1) m(0)
yreg(linear) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4 _IGenderRec_1
_IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5 _IFirstpsyc_6
_Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2 _Imaritalre_3
_Imaritalre_4)

drop if accepted_year_recode>=6
paramed noofadmissions_36m, avar(Cannabis1m) mvar(antipsy36m) a0(0) a1(1)
m(0) yreg(negbin) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
paramed mhasection36m, avar(Cannabis1m) mvar(antipsy36m) a0(0) a1(1) m(0)
yreg(logistic) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
paramed los_36m, avar(Cannabis1m) mvar(antipsy36m) a0(0) a1(1) m(0)
yreg(linear) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4 _IGenderRec_1
_IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5 _IFirstpsyc_6
_Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2 _Imaritalre_3
_Imaritalre_4)

drop if accepted_year_recode>=5
paramed noofadmissions_48m, avar(Cannabis1m) mvar(antipsy48m) a0(0) a1(1)
m(0) yreg(negbin) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
paramed mhasection48m, avar(Cannabis1m) mvar(antipsy48m) a0(0) a1(1) m(0)
yreg(logistic) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
```

```
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
paramed los_48m, avar(Cannabis1m) mvar(antipsy48m) a0(0) a1(1) m(0)
yreg(linear) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4 _IGenderRec_1
_IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5 _IFirstpsyc_6
_Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2 _Imaritalre_3
_Imaritalre_4)

drop if accepted_year_recode>=4
paramed noofadmissions_60m, avar(Cannabis1m) mvar(antipsy60m) a0(0) a1(1)
m(0) yreg(negbin) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
paramed mhasection60m, avar(Cannabis1m) mvar(antipsy60m) a0(0) a1(1) m(0)
yreg(logistic) mreg(linear) cvars(_Iagegp_2 _Iagegp_3 _Iagegp_4
_IGenderRec_1 _IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5
_IFirstpsyc_6 _Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2
_Imaritalre_3 _Imaritalre_4)
paramed los_60m, avar(Cannabis1m) mvar(antipsy60m) a0(0) a1(1) m(0)
yreg(linear) mreg(linear) cvars_Iagegp_2 _Iagegp_3 _Iagegp_4 _IGenderRec_1
_IFirstpsyc_2 _IFirstpsyc_3 _IFirstpsyc_4 _IFirstpsyc_5 _IFirstpsyc_6
_Iethnicrec_2 _Iethnicrec_3 _Iethnicrec_4 _Imaritalre_2 _Imaritalre_3
_Imaritalre_4)
```

The first set of commands (**Recoding demographic variables to deal with missing data and regrouping categories**) recodes numerical covariates to take into account missing covariate data. The second set of commands (**Label variables and create ordinal exposures**) labels categories for each variable. Descriptive statistics are provided for each of the covariates and outcome variables in relation to cannabis (**Descriptive statistics**). For outcome variables, if statements are used to select patients with sufficient follow-up data available for each outcome (“if accepted_year_recode<=X”). Multivariable analyses on each outcome are then performed (“**Multivariable analyses with missing data included**”). Employment status and accommodation status were not analysed or included in multivariable analyses owing to large amounts of missing data. The only covariate with missing data was marital status. As there were only 83 participants with missing marital status data, separate analyses were not performed to investigate the impact of dropping these cases from multivariable analyses.

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

The mediation analysis was performed by first examining the association of number of antipsychotics with each outcome (**Antipsychotic count predictor**) and then adding in the number of antipsychotics as a covariate to outcome analyses on cannabis (**Adding in antipsychoticcount as covariate**). The mediation analysis was performed using the paramed package in STATA (described in further detail in the main manuscript, section 4.1). Owing to constraints on the maximum length of variable names, shorter variable names were generated for antipsychotic count variables and dummy variables were created for each categorical covariate (**Generate shorter variable names and dummy variables**). Finally, the mediation analysis was performed (**Mediation analysis - paramed**). A sequential series of analyses from 1 year to 5 year follow-up was performed by iteratively dropping cases with shorter follow-up duration.

The output of the mediation analysis is as follows:

Natural direct effect: estimation of the association of cannabis use with clinical outcomes not mediated by number of antipsychotics, taking into account a baseline level of antipsychotic treatment for the population.

Natural indirect effect: estimation of the association of cannabis use with clinical outcomes which are potentially mediated by number of antipsychotics, taking into account a baseline level of antipsychotic treatment for the population.

Controlled direct effect: estimation of the association of cannabis use with clinical outcomes not mediated by number of antipsychotics, where the baseline level of the mediating variable (i.e. antipsychotic treatment) is fixed at a particular value.

Total effect: estimation of overall effect of cannabis use with clinical outcomes irrespective of the potential mediator.

The parameters for mediation analysis on each outcome were specified as follows:

avar: Cannabis1m – i.e. the predictor variable

4. Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis

mvar: antpsyXXX – i.e. the mediator variable

a0: 0 – baseline exposure level

a1: 1 – maximum exposure level

In this case, specifying a0=0 and a1=1 estimates mediation with a full range of possible exposure levels for the predictor variable. In cases where a population might always have a certain degree of exposure, the a0 value may be increased to reflect this (for example, if the exposure were background radiation or air pollution where the level of exposure can never be zero).

m: 0 – level of mediator for controlled direct effect

In this case, specifying m=0 estimates the controlled direct effect with the assumption that, by default, there is no exposure to the mediator (in this case that by default, patients received no antipsychotic medication).

yreg: negbin (i.e. negative binomial regression), logistic (i.e. binary logistic regression) or linear (i.e. linear regression) depending on the outcome variable.

mreg: linear – i.e. treating the antipsychotic count as a linear distribution

cvars: covariate dummy variables for the same covariates as analysed in the main analyses

5. Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method

This chapter incorporates an article published in BMJ Open investigating the association of negative symptoms with clinical outcomes in people with schizophrenia using CRIS and NLP. The article and accompanying supplementary data are presented in section 5.1. Further supplementary methods describing SQL data extraction and statistical analysis using STATA are described in section 5.2.

5.1 BMJ Open journal article

Please see overleaf.

BMJ Open

Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method

Rashmi Patel,[1] Nishamali Jayatilleke,[2] Matthew Broadbent,[3] Chin-Kuo Chang,[2] Nadia Foskett,[4] Genevieve Gorrell,[5] Richard D Hayes,[2] Richard Jackson,[2] Caroline Johnston,[6] Hitesh Shetty,[3] Angus Roberts,[5] Philip McGuire,[1] Robert Stewart[2]

To cite: Patel R, Jayatilleke N, Broadbent M, *et al*. Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method. *BMJ Open* 2015;**5**:e007619. doi:10.1136/bmjopen-2015-007619

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-007619).

RP, NJ, PM and RS contributed equally.

Received 8 January 2015
Revised 10 June 2015
Accepted 14 July 2015

For numbered affiliations see end of article.

Correspondence to
Dr Rashmi Patel;
bmj@rpatel.co.uk

ABSTRACT

Objectives: To identify negative symptoms in the clinical records of a large sample of patients with schizophrenia using natural language processing and assess their relationship with clinical outcomes.

Design: Observational study using an anonymised electronic health record case register.

Setting: South London and Maudsley NHS Trust (SLaM), a large provider of inpatient and community mental healthcare in the UK.

Participants: 7678 patients with schizophrenia receiving care during 2011.

Main outcome measures: Hospital admission, readmission and duration of admission.

Results: 10 different negative symptoms were ascertained with precision statistics above 0.80. 41% of patients had 2 or more negative symptoms. Negative symptoms were associated with younger age, male gender and single marital status, and with increased likelihood of hospital admission (OR 1.24, 95% CI 1.10 to 1.39), longer duration of admission (β-coefficient 20.5 days, 7.6–33.5), and increased likelihood of readmission following discharge (OR 1.58, 1.28 to 1.95).

Conclusions: Negative symptoms were common and associated with adverse clinical outcomes, consistent with evidence that these symptoms account for much of the disability associated with schizophrenia. Natural language processing provides a means of conducting research in large representative samples of patients, using data recorded during routine clinical practice.

Strengths and limitations of this study

- This is the largest known study (over 7000 participants) to investigate the relationship of negative symptoms with clinical outcomes in people with schizophrenia. Our findings demonstrate that negative symptoms are present in a substantial number of people with schizophrenia and are associated with increased hospital admission, readmission and duration of inpatient stay.
- To our knowledge, this is the first published study to use an automated information extraction method to acquire data on negative symptoms from electronic health records. This approach permits rapid acquisition of negative symptom data which is representative of everyday clinical practice in secondary mental healthcare.
- Our findings are based on data recorded by clinicians delivering routine mental healthcare who were not specifically ascertaining negative symptoms. It is therefore possible that negative symptoms were not comprehensively documented in the electronic health records from which they were identified leading to an inaccurate estimate of their prevalence in the analysed sample.

INTRODUCTION

Negative symptoms, which include amotivation, a flattening of emotional responses, a reduction in speech and activity, and social withdrawal,[1] contribute to much of the disability associated with schizophrenia.[2] These symptoms are also associated with poor psychosocial functioning[3] and a reduced likelihood of remission.[4–9] The aetiology and pathophysiology of negative symptoms are unknown, and there are no effective treatments.[10 11]

A number of excellent rating scales have been developed to assess negative symptoms.[12–14] However, these are relatively detailed, require a trained rater, and are not routinely applied in clinical practice. As a result, much of our knowledge of negative symptoms is derived from studies in relatively small samples of patients, who may have been selected for inclusion because they had particularly severe symptoms. The findings from these samples may not therefore be representative of negative symptoms in the overall population of patients with schizophrenia.

Clinical information is increasingly recorded electronically, facilitating access of rich clinical data, including presence or absence of symptoms,[15] from routine medical records. In the present study, we used a novel information extraction tool to identify negative symptomatology in a large body of electronic records collected from individuals with schizophrenia.[16–18] We then examined the relationship between negative symptoms and clinical outcomes. We tested the hypothesis that negative symptoms are common in patients with schizophrenia, and are associated with poor clinical outcome, as indexed by the frequency and duration of hospital admissions.

METHODS

Participants and clinical data

The study was carried out using the South London and Maudsley NHS Foundation Trust (SLaM) Biomedical Research Centre (BRC) Case Register, comprising electronic health record data dating back to April 2006 from a large mental healthcare provider to 1.2 million residents of southeast London (UK). The data were interrogated using the Clinical Record Interactive Search (CRIS) application,[19] with a robust anonymisation process and patient-led oversight.[20] Three samples were identified for analysis:

I. Sample A (n=7678): patients with schizophrenia (International Classification of Diseases (ICD)-10 F20.XX) aged 16 years and over who had used SLaM services during 2011. This sample was used to investigate: (1) the relationship between negative symptoms, documented at any point in the electronic health record, and demographic and other clinical measures (described below); (2) the relationship between negative symptoms documented prior to 1 January 2011 and the risk of hospital admission during 2011. This year was chosen for analysis because it maximised the duration of time over which text would be available for measurement development, while allowing at least 12 months follow-up in all instances.

II. Sample B (n=1612): the subset of patients from sample A who had been discharged from SLaM inpatient care during 2011. This sample was used to investigate the relationship between negative symptoms documented prior to 2011 and the risk of readmission in the 12 months following discharge.

III. Sample C (n=1609): the subset of patients from sample A who received SLaM inpatient care during 2011. This sample was used to investigate the relationship between negative symptoms documented prior to 2011 and the length of the first hospital admission during 2011.

Measurement development

Natural language processing (NLP) information extraction allows structured information to be obtained from unstructured text records. We used NLP to detect statements in the correspondence fields of clinical records to determine references to prespecified negative symptoms. Full details of the NLP method are described in a previous paper.[16] In summary, a putative training data set was selected which contained broad dictionary terms relevant to the negative symptoms of interest (described below). A detailed review of the training data set was undertaken by two psychiatrists (RP and RS) to identify and annotate key phrases within the records that were either relevant or irrelevant for keywords related to each symptom. Inter-rater reliability was tested between the two annotators resulting in percentage agreement of 93.0% (Cohen's κ 0.85). This training data set was used to construct an application (CRIS Negative Symptoms Scale, CRIS-NSS) using a hybrid classification model consisting of a support vector machine (SVM) learning algorithm[21] and rule-based text matching, using the Generalised Architecture for Text Engineering (GATE) software package.[17] The SVM algorithm was applied using a 'bag-of-words' approach to take into account the context of negative symptoms within the sentence in which they were documented, thereby allowing ascertainment of negative symptoms experienced specifically by the patient as well as distinguishing between positive instances and negated instances.[16] Once developed, CRIS-NSS was subsequently used to determine the presence of negative symptoms within the clinical sample. The accuracy of CRIS-NSS was evaluated using precision and recall statistics which were generated through internal fivefold cross-validation:[21] precision, representing the proportion of text instances identified by the tool which were found to be correct in terms of identifying the negative symptom of interest (equivalent to positive predictive value); and recall, measuring the proportion of text instances recording a given negative symptoms which were correctly identified as such by the tool (equivalent to sensitivity).

Details of the criteria for ascertaining the negative symptoms in the CRIS-NSS application are described in further detail elsewhere;[16] briefly, applications were developed for 10 items: poor motivation, blunted or flattened affect, poor eye contact, emotional withdrawal, poor rapport, social withdrawal, poverty of speech, mutism, apathy and concrete thinking. Each of these symptoms was defined as a binary variable on the basis of being present at any point in the record within the defined time period, and a composite scale (range 0–10) was constructed by summing these variables, followed by Cronbach α score calculation (a measure of intercorrelation between individual scale items) to estimate its internal consistency. A threshold score of at least 2 (ie, two or more negative symptoms documented) was applied a priori to determine the presence or absence of negative symptoms for analysis as a binary variable, as well as treating the scale score as an ordinal variable.

Clinical outcome measures and covariates

The following clinical and demographic variables were obtained as covariates from the data set: age (on 1 January 2011), gender, marital status, employment status, and admission and discharge dates for inpatient care episodes. Using structured data derived from the Health of the Nation Outcome Scale (HoNOS),[22] routinely completed in SLaM patients, the following subscales (scored 0–4) were used as covariates: activities of daily living (ADL) impairment, problems with relationships (social impairment), presence of hallucinations or delusions (a measure of positive symptoms) and depressive symptoms. For all of these HoNOS subscales, binary variables were defined on the basis of a score of 2 or more indicating the presence of each construct at levels judged to be clinically significant. In cases with multiple data points, all covariates were defined as those recorded closest to 1 January 2011.

Statistical analysis

STATA (V.11) software was used. Estimates of prevalence of negative symptoms by demographic factors were obtained as the proportion of patients within each group with two or more negative symptoms. After describing the distribution of negative symptoms and the psychometric properties of the CRIS-NSS, further analyses were performed to investigate the associations between the clinical outcomes described above and (1) the presence of negative symptoms, using binary logistic regression; and (2) CRIS-NSS scores, using ordinal logistic regression. Reference groups for categorical variables were generally defined as the most prevalent category, apart from age group where the youngest group of sufficient size was assigned as the reference. Associations between negative symptomatology and hospital admission and readmission were analysed using logistic regression, while those with length of inpatient stay were analysed using linear regression—again, estimating associations with both the binary and ordinal CRIS-NSS exposure. For the analyses with hospitalisation outcomes in/following 2011, CRIS-NSS was generated restricting information extraction to electronic health records prior to 2011. Where data were missing on individual covariates (in 2362 participants), this was indicated in the regression models as a separate category, supplemented by sensitivity analyses performed on the sample with complete data on all covariates to check the consistency of findings. A further supplementary analysis was performed to test the hypothesis that the association between negative symptoms and clinical outcomes varies with age. For this analysis, the previous analyses were repeated within the subgroups of those aged under the age of 40 years and those over the age of 40 years and including an interaction term of age under or over 40 and binary CRIS-NSS exposure. Finally, secondary analyses were undertaken to investigate and compare the relationships of individual CRIS-NSS symptoms with risk of readmission and length of stay using binary logistic and linear regression, respectively.

RESULTS

Performance of CRIS-NSS

Table 1 illustrates results from fivefold cross-validation of the CRIS-NSS tool. Precision coefficients ranged between 0.80 and 0.99 and recall between 0.62 and 0.97. For the composite 10-point scale, the Cronbach α value was 0.78 indicating a good level of internal consistency.

Prevalence and distribution of negative symptoms

Of the 7678 patients with schizophrenia, 3149 (41.0%) had at least two negative symptoms documented. Table 1 displays prevalences for each of the symptoms classified by the tool. The most frequently recorded symptoms were poor motivation (30.5%), blunted or flattened affect (27.4%), poor eye contact (26.0%) and emotional withdrawal (23.5%). The prevalences by number of symptoms were as follows: one symptom 14.6%, two symptoms 12.7%, three symptoms 9.3%, four symptoms 6.4%, five symptoms 5.0%, six or more symptoms 7.6%.

Binary logistic regression analyses (table 2) revealed that patients with two or more negative symptoms were most likely to be 20–29 years old, male and single. Two or more negative symptoms were also associated with ADL impairment, whereas patients who were employed were less likely to have negative symptoms compared with those unemployed. Ordinal logistic regression analysis (etable 1) revealed similar findings for CRIS-NSS score as an exposure, and sensitivity analyses limited to those with full data on all covariates (etable 2) were also consistent.

Table 1 Performance of Clinical Record Interactive Search Negative Symptoms Scale (CRIS-NSS) information extraction applications ascertaining individual symptom domains

Symptom	Precision/recall	Prevalence (%) in patients with schizophrenia receiving care during 2011 (n=7678)
Poor motivation	0.87/0.62	30.5
Blunted or flattened affect	0.93/0.83	27.4
Poor eye contact	0.95/0.79	26.0
Emotional withdrawal	0.85/0.74	23.5
Poor rapport	0.91/0.77	16.3
Social withdrawal	0.94/0.96	12.7
Poverty of speech	0.80/0.73	12.4
Mute	0.99/0.94	8.1
Apathy	0.88/0.97	7.7
Concrete thinking	0.91/0.72	5.7

Table 2 Binary logistic regression analysis of factors associated with negative symptoms in patients with schizophrenia (n=7678)

Factor	Group	Number in sample	Prevalence of two or more negative symptoms (%)	Association with two or more negative symptoms: OR (95% CI), p value			
				Unadjusted		Adjusted model (n=7676)*	
Age (years)	16–19	203	27.6	0.35 (0.25 to 0.49)	<0.001	0.50 (0.35 to 0.71)	<0.001
	20–29	1337	52.0	Reference		Reference	
	30–39	1775	47.0	0.82 (0.71 to 0.94)	0.006	0.85 (0.73 to 0.99)	0.033
	40–49	1983	42.6	0.69 (0.60 to 0.79)	<0.001	0.71 (0.61 to 0.82)	<0.001
	50–59	1137	37.2	0.55 (0.47 to 0.64)	<0.001	0.56 (0.47 to 0.67)	<0.001
	60–69	654	29.1	0.38 (0.31 to 0.46)	<0.001	0.39 (0.31 to 0.48)	<0.001
	70+	589	18.0	0.20 (0.16 to 0.26)	<0.001	0.22 (0.17 to 0.28)	<0.001
Gender	Male	4592	45.3	Reference		Reference	
	Female	3084	34.7	0.64 (0.59 to 0.71)	<0.001	0.77 (0.70 to 0.85)	<0.001
Marital status (most recent)	Single	5795	44.6	Reference		Reference	
	Married/cohabiting	785	31.6	0.57 (0.49 to 0.67)	<0.001	0.76 (0.64 to 0.90)	0.002
	Divorced/separated	776	33.4	0.62 (0.53 to 0.73)	<0.001	0.85 (0.71 to 1.00)	0.054
	Widowed	208	21.2	0.33 (0.24 to 0.47)	<0.001	0.77 (0.53 to 1.12)	0.178
Employment (most recent)	Unemployed	4956	47.9	Reference		Reference	
	Employed	341	39.6	0.71 (0.57 to 0.89)	0.003	0.68 (0.54 to 0.86)	0.001
	In education	311	39.6	0.71 (0.56 to 0.90)	0.004	0.81 (0.63 to 1.03)	0.089
	Retired	7	14.3	0.18 (0.02 to 1.51)	0.114	0.40 (0.04 to 3.52)	0.405
ADL impairment	Absent	4700	41.9	Reference		Reference	
	Present	2283	46.3	1.20 (1.08 to 1.32)	<0.001	1.35 (1.21 to 1.52)	<0.001
Social impairment	Absent	4432	42.7	Reference		Reference	
	Present	2533	44.4	1.07 (0.97 to 1.18)	0.158	0.94 (0.84 to 1.05)	0.240
Delusions/hallucinations	Absent	3904	41.9	Reference		Reference	
	Present	3077	45.0	1.14 (1.03 to 1.25)	0.009	1.19 (1.07 to 1.31)	0.001
Depression	Absent	4976	45.2	Reference		Reference	
	Present	2014	38.8	0.77 (0.69 to 0.85)	<0.001	0.74 (0.66 to 0.82)	<0.001

*Results adjusted for all the factors reported in this table; two cases with no recorded data on gender were dropped.
ADL, activities of daily living.

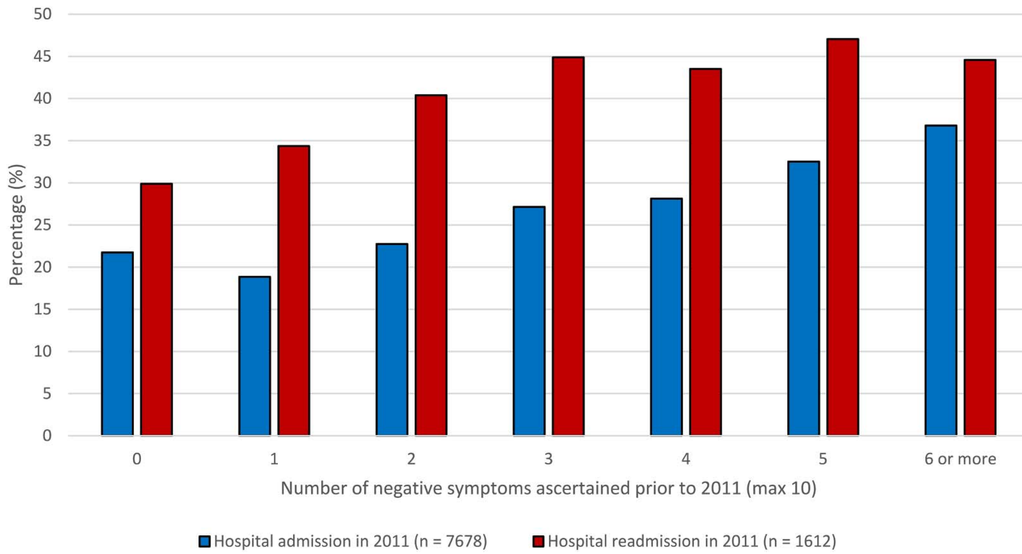


Figure 1 Percentage of patients admitted to hospital or readmitted to hospital following discharge in 2011 by number of negative symptoms.

Hospital admission, length of stay and readmission

Figure 1 summarises the association of negative symptoms recorded prior to 2011 with mental health admission (etable 3) and readmission (etable 4) in 2011. Figure 2 summarises length of hospitalisation for inpatients during 2011 (etable 5). Logistic and linear regression analyses (table 3) confirmed that negative symptoms were associated with a higher likelihood of admission, readmission and a longer duration of hospitalisation. Specifically, after full adjustment (table 3, model 3), patients with two or more negative symptoms before 2011 had a 24% greater likelihood of admission during 2011. Moreover, each of their admissions was, on average, an extra 21 days in duration, and when they were discharged, they had a 58% higher risk of readmission within 12 months. All of these associations remained independent and largely unaltered following adjustment for intensity of delusions/hallucinations among other covariates. Further analysis (etable 6) comparing patients aged under and over 40 years showed that the effects of negative symptoms on inpatient admission were broadly similar for both groups but with a slight increase in risk of readmission and reduced duration of admission in relation to negative symptoms for those under 40 compared with those over 40. However, the age × negative symptoms interaction term remained a non-significant factor ($p>0.05$) for all models.

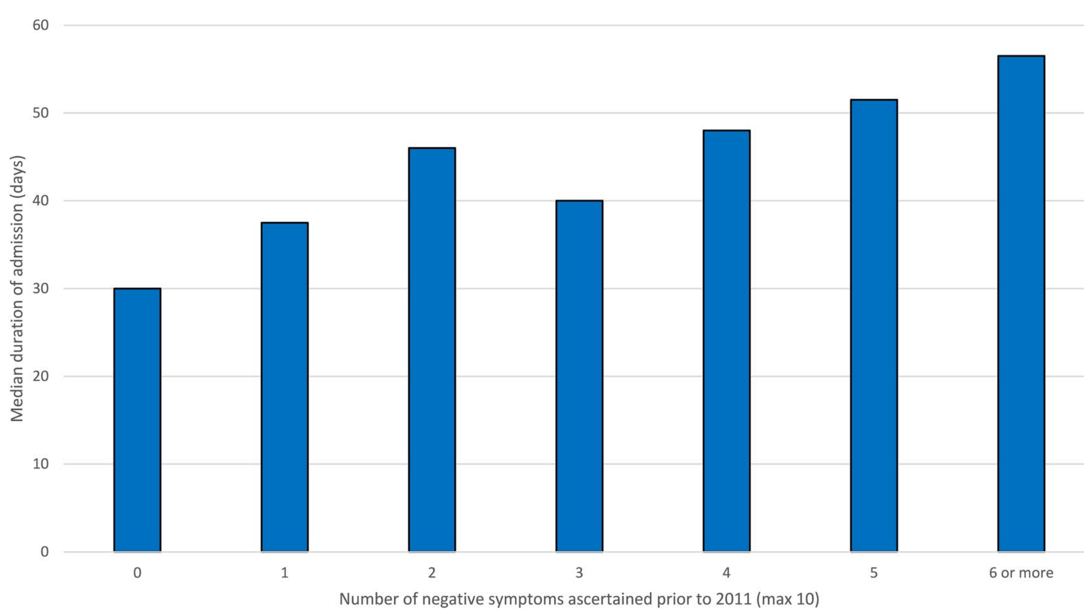


Figure 2 Median duration of admission among mental health inpatients with schizophrenia in 2011 by number of negative symptoms (n=1609).

Table 3 Association between number of negative symptoms ascertained prior to 2011 and mental health hospital admission, readmission and duration of admission in 2011

	Inpatient admission (OR, 95% CI; n=7678)*	Readmission within 12 months of inpatient admission (OR, 95% CI; n=1612)*	Duration of inpatient admission (days; β-coefficient, 95% CI; n=1609)†
Associations with 2 or more negative symptoms (binary variable)			
Unadjusted	1.47 (1.32 to 1.63)	1.73 (1.41 to 2.12)	23.9 (11.2 to 36.7)
1. Age and sex	1.37 (1.23 to 1.53)	1.70 (1.38 to 2.09)	24.1 (11.3 to 36.9)
2. Model 1 plus marital status and employment	1.27 (1.13 to 1.42)	1.58 (1.28 to 1.96)	20.1 (7.1 to 33.1)
3. Model 2 plus delusions/hallucinations, and depression	1.24 (1.10 to 1.39)	1.58 (1.28 to 1.95)	20.5 (7.6 to 33.5)
Associations with incremental number of negative symptoms (10-point scale ordinal variable)‡			
Unadjusted	1.12 (1.09 to 1.15)	1.12 (1.07 to 1.17)	6.5 (3.5 to 9.4)
1. Age and sex	1.09 (1.06 to 1.12)	1.11 (1.06 to 1.16)	6.3 (3.3 to 9.2)
2. Model 1 plus marital status and employment	1.07 (1.04 to 1.10)	1.09 (1.04 to 1.14)	5.4 (2.4 to 8.4)
3. Model 2 plus delusions/hallucinations, and depression	1.07 (1.04 to 1.10)	1.09 (1.04 to 1.14)	5.6 (2.6 to 8.6)

*Logistic regression.
†Linear regression.
‡ORs and β-coefficients are per one unit increase on the 10-point scale.

Finally, logistic and linear regression analyses were performed to examine the relationship between individual negative symptoms and the frequency and duration of admission (table 4). Poor eye contact and poor rapport were associated with increased risk of readmission, while apathy was associated with increased duration of admission. Emotional withdrawal and mutism were associated with both the risk of readmission and the duration of admission.

DISCUSSION

Using an SVM learning method with an NLP tool, we were able successfully to extract data on negative symptoms from the electronic mental health records of a large clinical sample of patients with schizophrenia. This approach did not require any specialised training or extra clinical assessments, and was able to generate a scale with robust construct and predictive validity from data recorded as part of routine clinical care.

The results suggest that negative symptoms are documented in the electronic health records of a sizeable proportion of patients with schizophrenia, particularly in those who are relatively young, male and not in a relationship, consistent with findings from studies that assessed negative symptoms using quite different methods.[23 24] Our findings were based on the unprompted documentation of negative symptoms in the context of routine clinical care by staff who were not specifically trained in their assessment. Previous findings have usually been based on systematic ratings by a researcher using a dedicated rating scale. Negative symptoms are relatively difficult to detect and assess,[1 2] and may be less frequently documented than positive symptoms, such as delusions and hallucinations, because they are less clinically obvious. In addition, mental health services in the UK are often orientated towards the management of acute crises, and hence the treatment of positive symptoms.[25] It is thus possible that the figures for the prevalence and the severity of negative symptoms derived from our approach are lower than would have been obtained from a trained assessor using a standardised instrument. In addition, our method may be more likely to identify the types of negative symptoms (eg, poverty of speech) whose detection does not require specialised training.

We found that a substantial proportion (41%) of the sample had at least two negative symptoms. Although we defined and assessed negative symptoms in different ways to previous studies, this figure is comparable to that described in other samples of patients with schizophrenia (Jager *et al*[4]: 44%; Bobes *et al*[23]: 58%; Cohen *et al*[24]: 40%). Taken together, these findings suggest that negative symptoms are a relatively common feature of schizophrenia, rather than being limited to a subgroup of patients with a chronic, unremitting illness.[26]

As predicted, we found a clear association between negative symptoms and poor clinical outcomes, as indexed by impairments in daily living, increased risk of admission, increased duration of admission and increased risk of readmission. Hospital admissions are the main drivers of cost in the care of patients with schizophrenia,[27] but have traditionally been linked to the severity of positive psychotic symptoms.[28] Our data indicate that negative symptoms are an equally important factor, and suggest that a greater emphasis on assessing and treating these features of schizophrenia may have significant health economic benefits. However, as our findings are drawn from observational data, it would be necessary to perform interventional clinical studies to determine whether an effective treatment for negative symptoms would lead to better clinical outcomes.

Table 4 Associations between individual Clinical Record Interactive Search Negative Symptoms Scale (CRIS-NSS) components and readmission risk/duration of admission in 2011

	Readmission risk (binary logistic regression) (n=1612)				Duration of admission (linear regression) (n=1590)			
	Unadjusted		Adjusted*		Unadjusted		Adjusted*	
Negative symptom	OR (95% CI)	p Value	OR (95% CI)	p Value	β-coefficient (95% CI)	p Value	β-coefficient (95% CI)	p Value
Poor motivation	1.40 (1.13 to 1.74)	0.002	1.29 (1.03 to 1.61)	0.026	23.0 (9.1 to 36.9)	0.001	19.1 (5.0 to 33.2)	0.008
Blunted or flattened affect	1.34 (1.08 to 1.65)	0.007	1.20 (0.97 to 1.50)	0.097	12.8 (−1.2 to 26.8)	0.073	8.3 (−5.7 to 22.4)	0.242
Poor eye contact	1.60 (1.30 to 1.98)	<0.001	1.48 (1.19 to 1.83)	<0.001	18.0 (4.2 to 31.8)	0.011	14.8 (0.9 to 28.6)	0.036
Emotional withdrawal	1.62 (1.30 to 2.02)	<0.001	1.49 (1.19 to 1.87)	0.001	32.5 (18.1 to 46.9)	<0.001	30.0 (15.6 to 44.4)	<0.001
Poor rapport	1.63 (1.29 to 2.06)	<0.001	1.50 (1.18 to 1.90)	0.001	23.1 (7.5 to 38.6)	0.004	21.1 (5.5 to 36.6)	0.008
Social withdrawal	1.16 (0.88 to 1.54)	0.291	1.02 (0.76 to 1.36)	0.911	16.4 (−2.9 to 35.7)	0.095	9.2 (−10.1 to 28.6)	0.349
Poverty of speech	1.30 (0.98 to 1.70)	0.064	1.12 (0.85 to 1.49)	0.416	13.2 (−5.8 to 32.2)	0.173	8.5 (−10.5 to 27.5)	0.379
Mute	1.71 (1.27 to 2.30)	<0.001	1.56 (1.15 to 2.12)	0.004	28.5 (7.9 to 49.1)	0.007	29.2 (8.6 to 49.7)	0.005
Apathy	1.02 (0.71 to 1.47)	0.914	0.93 (0.64 to 1.35)	0.692	32.5 (6.7 to 58.2)	0.013	27.4 (1.8 to 53.1)	0.036
Concrete thinking	1.37 (0.94 to 2.01)	0.100	1.25 (0.85 to 1.84)	0.250	16.8 (−10.2 to 43.9)	0.222	11.3 (−15.5 to 38.1)	0.407

*Adjusted for age, sex, marital status, employment status, presence of hallucinations/delusions and depression.

A key strength of our study was the large size of the patient sample, and that it was representative of the overall clinical population of people with schizophrenia in a defined geographic area. Previous studies of negative symptoms have usually involved smaller patient samples that were recruited to a research project.[4 23 24] Focusing the information extraction process on text from correspondence maximises the generalisability of our approach, as letters to primary care physicians (which accounted for a large portion of the correspondence text) are unlikely to vary substantially between mental health services with respect to the language used to describe the symptoms of interest. In the present study, we examined the patient's entire record rather than discrete periods of illness, and it was not possible to delineate the timing or duration of individual negative symptoms, or whether they were primary (ie, a direct consequence of illness) or secondary (eg, side effects of treatment) as these measures were not routinely documented in electronic health records. Although we investigated the association of negative symptoms in clinical documents prior to 1 January 2011 with outcomes occurring after 1 January 2011 (to ensure that negative symptoms were always ascertained prior to outcomes), if negative symptoms were identified prior to 1 January 2011, it was not possible to ascertain when they occurred prior to this date, or their temporal relationships to subsequent clinical outcomes. The findings were thus derived from assessments made over a period that was not standardised, but was generally relatively long. In contrast, most assessments of negative symptoms in the literature are derived from a single cross-sectional measurement.[29 30]

A further limitation of our analysis was the extent to which individual negative symptoms could be considered as having equal weight in a composite score. Weighting the 10 negative symptom applications equally resulted in a composite score (from 0 to 10) with a reasonable degree of internal consistency, as demonstrated by a Cronbach α value of 0.78. However, analysing the association of each negative symptom with clinical outcomes revealed varying degrees of association with poor clinical outcomes for different negative symptoms. Future studies are necessary to examine the propensity for different negative symptoms to co-occur in individual patients and the extent to which different clusters of symptoms are associated with clinical outcomes, particularly in the light of previous research which suggests that negative symptoms segregate into two subdomains relating to amotivation and reduced emotional expression.[31]

The application of NLP to clinical records is unlikely to identify negative symptoms as accurately as a direct assessment using a specialised psychopathological rating scale. However, automated tools could be used to screen individuals and identify those with negative symptoms who would then benefit from comprehensive assessment using a standardised instrument. In this way, automated methods could be used to complement standardised instruments. Automated information extraction tools could also be developed to identify other clinical parameters from electronic health records in order to support real-time clinical decision-making. These possibilities could be explored in future research.

In summary, our data suggest that negative symptoms can be identified in clinical records using automated methods, are common in patients with schizophrenia and are associated with poor clinical outcomes. The findings highlight the potential of automated information extraction tools in mental health research and clinical practice, and the importance of developing effective treatments for negative symptoms.

Author affiliations
[1]Department of Psychosis Studies, King's College London, Institute of Psychiatry, Psychology & Neuroscience, London, UK
[2]Department of Psychological Medicine, King's College London, Institute of Psychiatry, Psychology & Neuroscience, London, UK
[3]South London and Maudsley NHS Foundation Trust, Biomedical Research Centre Nucleus, London, UK
[4]Roche Products Limited, Welwyn Garden City, UK
[5]Department of Computer Science, The University of Sheffield, Portobello, Sheffield, UK
[6]Social Developmental and Genetic Psychiatry Department, King's College London, Institute of Psychiatry, Psychology & Neuroscience, London, UK

Contributors The study was conceived by RS and NF. The CRIS-NSS product development was led by RJ with significant input from MB, GG, CJ, AR and HS. Initial analyses were carried out by RS, C-KC and RDH. Final analyses and reporting of findings were led by RP and NJ, supervised by RS and PM. All authors contributed to manuscript preparation and approved the final version.

Funding NJ, MB, C-KC, RDH, CJ, RJ, HS and RS are funded by the National Institute for Health Research (NIHR) Biomedical Research Centre and Dementia Biomedical Research Unit at South London and Maudsley NHS Foundation Trust and King's College London, which also supports the development and maintenance of the CRIS data resource. The analyses reported here were part-funded by Roche. RDH is supported by a UK Medical Research Council Population Health Scientist Fellowship (MR/J01219X/1). RP is supported by a UK Medical Research Council Clinical Research Training Fellowship (MR/K002813/1).

Disclaimer Funding organisations had no role in the collection, management, analysis, and interpretation of the data; and the preparation, review, or approval of the manuscript.

Competing interests The CRIS team (MB, C-KC, RDH, RJ, HS and RS) have received research funding from Roche; Pfizer; Johnson and Johnson; and Lundbeck. PM has received research funding from Janssen; Sunovion; GW Pharmaceuticals; and Roche.

Ethics approval The CRIS data resource received ethical approval as an anonymised data set for secondary analyses from Oxfordshire REC C (Ref: 08/H0606/71+5).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Foussias G, Agid O, Remington G. *Handbook of schizophrenia spectrum disorders, volume II*. Dordrecht: Springer Netherlands, 2011.
2. Hunter R, Barry S. Negative symptoms and psychosocial functioning in schizophrenia: neglected but important targets for treatment. *Eur Psychiatry* 2012;27:432–6.
3. Rabinowitz J, Berardo CG, Bugarski-Kirola D, *et al*. Association of prominent positive and prominent negative symptoms and functional health, well-being, healthcare-related quality of life and family burden: A CATIE analysis. *Schizophr Res* 2013;150:339–42.
4. Jager M, Riedel M, Schmauss M, *et al*. Prediction of symptom remission in schizophrenia during inpatient treatment. *World J Biol Psychiatry* 2009;10:426–34.
5. Ucok A, Serbest S, Kandemir PE. Remission after first-episode schizophrenia: results of a long-term follow-up. *Psychiatry Res* 2011;189:33–7.
6. Moller HJ, Bottlender R, Wegner U, *et al*. Long-term course of schizophrenic, affective and schizoaffective psychosis: focus on negative symptoms and their impact on global indicators of outcome. *Acta Psychiatr Scand Suppl* 2000;407:54–7.
7. Dominguez M-G, Saka MC, can Saka M, *et al*. Early expression of negative/disorganized symptoms predicting psychotic experiences and subsequent clinical psychosis: a 10-year study. *Am J Psychiatry* 2010;167:1075–82.
8. McGurk SR, Moriarty PJ, Harvey PD, *et al*. Relationship of cognitive functioning, adaptive life skills, and negative symptom severity in poor-outcome geriatric schizophrenia patients. *J Neuropsychiatry Clin Neurosci* 2000;12:257–64.
9. Kirkpatrick B, Fenton WS, Carpenter WT, *et al*. The NIMH-MATRICS consensus statement on negative symptoms. *Schizophr Bull* 2006;32:214–19.
10. Arango C, Garibaldi G, Marder SR. Pharmacological approaches to treating negative symptoms: a review of clinical trials. *Schizophr Res* 2013;150:346–52.
11. Fusar-Poli P, Papanastasiou E, Stahl D, *et al*. Treatments of negative symptoms in schizophrenia: meta-analysis of 168 randomized placebo-controlled trials. *Schizophr Bull* 2014;41:892–9.
12. Blanchard JJ, Kring AM, Horan WP, *et al*. Toward the next generation of negative symptom assessments: the collaboration to advance negative symptom assessment in schizophrenia. *Schizophr Bull* 2011;37:291–9.
13. Kirkpatrick B, Strauss GP, Nguyen L, *et al*. The brief negative symptom scale: psychometric properties. *Schizophr Bull* 2011;37:300–5.
14. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull* 1987;13:261–76.
15. Patel R, Lloyd T, Jackson R, *et al*. Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. *BMJ Open* 2015;5:e007504.
16. Gorrell G, Jackson R, Roberts A, *et al*. Finding negative symptoms of schizophrenia in patient records. *Proceedings of NLP Med Biol Work (NLPMedBio), Recent Adv Nat Lang Process (RANLP), Hissar, Bulg* 2013:9–17. http://aclweb.org/anthology/W/W13/W13-5102.pdf
17. Cunningham H, Tablan V, Roberts A, *et al*. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput Biol* 2013;9:e1002854.
18. Patel R, Jayatilleke N, Jackson R, *et al*. Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach. *Lancet* 2014;383:S16.
19. Stewart R, Soremekun M, Perera G, *et al*. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* 2009;9:51.
20. Fernandes AC, Cloete D, Broadbent MTM, *et al*. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Med Inform Decis Mak* 2013;13:71.
21. Li Y, Bontcheva K, Cunningham H. Adapting SVM for data sparseness and imbalance: a case study in information extraction. *Nat Lang Eng* 2009;15:241–71.
22. Wing JK, Beevor AS, Curtis RH, *et al*. Health of the Nation Outcome Scales (HoNOS). Research and development. *Br J Psychiatry* 1998;172:11–18.
23. Bobes J, Arango C, Garcia-Garcia M, *et al*. Prevalence of negative symptoms in outpatients with schizophrenia spectrum disorders treated with antipsychotics in routine clinical practice: findings from the CLAMORS study. *J Clin Psychiatry* 2010;71:280–6.
24. Cohen CI, Natarajan N, Araujo M, *et al*. Prevalence of negative symptoms and associated factors in older adults with schizophrenia spectrum disorder. *Am J Geriatr Psychiatry* 2013;21:100–7.
25. Jacobs R, Barrenho E. Impact of crisis resolution and home treatment teams on psychiatric admissions in England. *Br J Psychiatry* 2011;199:71–6.

26. Bagney A, Rodriguez-Jimenez R, Martinez-Gras I, *et al*. Negative symptoms and executive function in schizophrenia: does their relationship change with illness duration? *Psychopathology* 2013;46:241–8.
27. Department of Health. National survey of Investment in Mental Health Services. 2013:1596–11. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/140098/FinMap2012-NatReportAdult-0308212.pdf
28. Olfson M, Ascher-Svanum H, Faries DE, *et al*. Predicting psychiatric hospital admission among adults with schizophrenia. *Psychiatr Serv* 2011;62:1138–45.
29. Gilbert EA, Liberman RP, Ventura J, *et al*. Concurrent validity of negative symptom assessments in treatment refractory schizophrenia: relationship between interview-based ratings and inpatient ward observations. *J Psychiatr Res* 2000;34:443–7.
30. Chang WC, Hui CLM, Tang JYM, *et al*. Persistent negative symptoms in first-episode schizophrenia: a prospective three-year follow-up study. *Schizophr Res* 2011;133:22–8.
31. Kimhy D, Yale S, Goetz RR, *et al*. The factorial structure of the schedule for the deficit syndrome in schizophrenia. *Schizophr Bull* 2006;32:274–8.

Title: Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method

Authors: Rashmi Patel, BM, BCh[1a]; Nishamali Jayatilleke, BM, MPhil[2a]; Matthew Broadbent, BSc[3]; Chin-Kuo Chang, PhD[2]; Nadia Foskett, PhD[4]; Genevieve Gorrell DPhil[5]; Richard D Hayes, PhD[2]; Richard Jackson, MSc[2]; Caroline Johnston, PhD[6]; Hitesh Shetty, MSc[3]; Angus Roberts PhD[5]; Philip McGuire, MD, PhD[1b]; Robert Stewart, MD[2b]

[a,b]Equal contributions to authorship

Author affiliations:

1. King's College London, Department of Psychosis Studies, Institute of Psychiatry, Box PO 63, De Crespigny Park, Denmark Hill, London SE5 8AF, UK

2. King's College London, Department of Psychological Medicine, Institute of Psychiatry, Box PO 92, De Crespigny Park, Denmark Hill, London SE5 8AF, UK

3. South London and Maudsley NHS Foundation Trust, Biomedical Research Centre Nucleus, Mapother House, De Crespigny Park, Denmark Hill, London SE5 8AF, UK

4. Roche Products Limited, 6 Falcon Way, Shire Park, Welwyn Garden City AL7 1TW, UK

5. The University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello, Sheffield S1 4DP, UK

6. King's College London, Social Developmental and Genetic Psychiatry Department, Institute of Psychiatry, Box PO 80, De Crespigny Park, Denmark Hill, London SE5 8AF, UK

Correspondence to:

Rashmi Patel BM BCh, MRC Clinical Research Training Fellow, King's College London, Department of Psychosis Studies, Institute of Psychiatry, Box PO 63, De Crespigny Park, Denmark Hill, London SE5 8AF, UK

Telephone: +44 2078480355

E-mail: bmj@rpatel.co.uk

Supplementary material

eTable 1 - Multivariable ordinal logistic regression analysis of factors associated with negative symptoms in patients with schizophrenia								
Factor	**Group**	**Number in sample**	**CRIS-NSS mean score (range 0-10)**	**Standard Deviation**	**Unadjusted**		**Adjusted model (n=7676)***	
					Odds ratio (95% CI)	**P value**	**Odds ratio (95% CI)**	**P value**
Age (years)	16-19	203	1.28	2.15	0.30 (0.22-0.40)	<0.001	0.46 (0.33-0.64)	<0.001
	20-29	1337	2.45	2.64	Reference		Reference	
	30-39	1775	1.97	2.21	0.77 (0.67-0.87)	<0.001	0.78 (0.68-0.89)	<0.001
	40-49	1983	1.66	1.95	0.62 (0.55-0.71)	<0.001	0.62 (0.55-0.71)	<0.001
	50-59	1137	1.43	1.79	0.53 (0.46-0.61)	<0.001	0.52 (0.45-0.61)	<0.001
	60-69	654	1.07	1.53	0.37 (0.31-0.44)	<0.001	0.36 (0.30-0.44)	<0.001
	70+	589	0.72	1.19	0.25 (0.21-0.30)	<0.001	0.25 (0.20-0.31)	<0.001
Gender	Male	4592	1.90	2.22	Reference		Reference	
	Female	3084	1.41	1.91	0.66 (0.60-0.71)	<0.001	0.79 (0.72-0.86)	<0.001
Marital status (most recent)	Single	5795	1.88	2.20	Reference		Reference	
	Married/cohabiting	785	1.16	1.61	0.57 (0.50-0.66)	<0.001	0.74 (0.64-0.86)	<0.001
	Divorced/separated	776	1.35	1.86	0.64 (0.56-0.74)	<0.001	0.86 (0.74-0.99)	0.040
	Widowed	208	0.85	1.39	0.40 (0.31-0.52)	<0.001	0.81 (0.60-1.09)	0.160
Employment (most recent)	Unemployed	4956	2.03	2.23	Reference		Reference	
	Employed	341	1.51	1.87	0.68 (0.56-0.83)	<0.001	0.64 (0.52-0.78)	<0.001
	In education	311	1.64	2.10	0.70 (0.57-0.86)	0.001	0.78 (0.63-0.97)	0.026
	Retired	7	0.57	0.79	0.33 (0.08-1.29)	0.110	0.68 (0.16-2.87)	0.599
ADL impairment	Absent	4700	1.73	2.09	Reference		Reference	
	Present	2283	1.97	2.23	1.21 (1.11-1.33)	<0.001	1.35 (1.22-1.49)	<0.001
Social impairment	Absent	4432	1.76	2.10	Reference		Reference	
	Present	2533	1.88	2.20	1.07 (0.98-1.17)	0.108	0.95 (0.86-1.05)	0.292
Delusions / hallucinations	Absent	3904	1.77	2.17	Reference		Reference	
	Present	3077	1.85	2.11	1.14 (1.05-1.24)	0.003	1.19 (1.09-1.30)	<0.001
Depression	Absent	4976	1.90	2.16	Reference		Reference	
	Present	2014	1.59	2.08	0.71 (0.65-0.79)	<0.001	0.69 (0.62-0.76)	<0.001

*Results adjusted for all the factors reported in this table; 2 cases with no recorded data on gender were dropped.

eTable 2 - Multivariable binary logistic regression analysis of factors associated with negative symptoms in patients with schizophrenia including cases with full covariate data only

Factor	Group	Number in sample	Prevalence of two or more negative symptoms (%)	Association with two or more negative symptoms: odds ratio (95% CI), p-value			
				Unadjusted		Adjusted model (n=5316)*	
Age (years)	16-19	203	27.6	0.35 (0.25-0.49)	<0.001	0.44 (0.29-0.68)	<0.001
	20-29	1337	52.0	Reference		Reference	
	30-39	1775	47.0	0.82 (0.71-0.94)	0.006	0.80 (0.67-0.95)	0.012
	40-49	1983	42.6	0.69 (0.60-0.79)	<0.001	0.64 (0.54-0.76)	<0.001
	50-59	1137	37.2	0.55 (0.47-0.64)	<0.001	0.45 (0.37-0.55)	<0.001
	60-69	654	29.1	0.38 (0.31-0.46)	<0.001	0.32 (0.25-0.41)	<0.001
	70+	589	18.0	0.20 (0.16-0.26)	<0.001	0.13 (0.09-0.18)	<0.001
Gender	Male	4592	45.3	Reference		Reference	
	Female	3083	34.7	0.64 (0.59-0.71)	<0.001	0.74 (0.65-0.83)	<0.001
Marital status (most recent)	Single	5795	44.6	Reference		Reference	
	Married/cohabiting	785	31.6	0.57 (0.49-0.67)	<0.001	0.77 (0.63-0.94)	0.012
	Divorced/separated	776	33.4	0.62 (0.53-0.73)	<0.001	0.91 (0.75-1.12)	0.376
	Widowed	208	21.2	0.33 (0.24-0.47)	<0.001	0.85 (0.54-1.32)	0.466
Employment (most recent)	Unemployed	4956	47.9	Reference		Reference	
	Employed	341	39.6	0.71 (0.57-0.89)	0.003	0.65 (0.51-0.83)	<0.001
	In education	311	39.6	0.71 (0.56-0.90)	0.004	0.78 (0.61-1.02)	0.065
	Retired	7	14.3	0.18 (0.02-1.51)	0.114	0.51 (0.06-4.65)	0.547
ADL impairment	Absent	4700	41.9	Reference		Reference	
	Present	2283	46.3	1.20 (1.08-1.32)	<0.001	1.29 (1.13-1.47)	<0.001
Social impairment	Absent	4432	42.7	Reference		Reference	
	Present	2533	44.4	1.07 (0.97-1.18)	0.158	0.93 (0.82-1.05)	0.258
Delusions / hallucinations	Absent	3904	41.9	Reference		Reference	
	Present	3077	45.0	1.14 (1.03-1.25)	0.009	1.23 (1.10-1.38)	<0.001
Depression	Absent	4976	45.2	Reference		Reference	
	Present	2014	38.8	0.77 (0.69-0.85)	<0.001	0.69 (0.61-0.78)	<0.001

*Results adjusted for all the factors reported in this table.

eTable 3 - Percentage of patients admitted to hospital in 2011 by number of negative symptoms (n=7678)		
Number of negative symptoms	**Number of patients**	**Percentage admitted to hospital in 2011 (%)**
0	3408	21.7
1	1121	18.9
2	974	22.7
3	717	27.2
4	492	28.1
5	382	32.5
6 or more	584	36.8

eTable 4 - Percentage of patients readmitted to hospital following discharge in 2011 by number of negative symptoms (n=1612)		
Number of negative symptoms	**Number of patients**	**Percentage admitted to hospital in 2011 (%)**
0	612	29.9
1	195	34.4
2	213	40.4
3	176	44.9
4	131	43.5
5	119	47.1
6 or more	166	44.6

eTable 5 - Median duration of admission amongst mental health inpatients with schizophrenia in 2011 by number of negative symptoms (n=1,609)		
Number of negative symptoms	**Number of patients**	**Median duration of admission (days)**
0	696	30.0
1	200	37.5
2	194	46.0
3	165	40.0
4	116	48.0
5	110	51.5
6 or more	128	56.5

eTable 6 - Association between number of negative symptoms ascertained prior to 2011 and mental health hospital admission, re-admission and duration of admission in 2011 in patients aged under 40 years and patients aged over 40 years			
	Inpatient admission (odds ratio, 95% CI)*	**Re-admission within 12 months of inpatient admission (odds ratio, 95% CI)***	**Duration of inpatient admission (days; B-coefficient, 95% CI)****
Associations with 2 or more negative symptoms (binary variable) in patients aged between 16 and 39 years.	n=3315	n=792	n=785
Unadjusted	1.36 (1.17-1.59)	1.82 (1.36-2.43)	25.4 (6.2, 44.6)
1. Age and sex	1.40 (1.20-1.63)	1.88 (1.40-2.54)	20.8 (1.5, 40.1)
2. Model 1 plus marital status and employment	1.25 (1.06-1.46)	1.70 (1.24-2.31)	15.0 (-4.9, 34.9)
3. Model 2 plus delusions / hallucinations, and depression	1.22 (1.03-1.43)	1.68 (1.23-2.29)	14.5 (-5.5, 34.5)
Associations with 2 or more negative symptoms (binary variable) in patients aged over 40 years.	n=4361	n=820	n=805
Unadjusted	1.41 (1.20-1.65)	1.61 (1.21-2.16)	22.1 (5.2, 39.1)
1. Age and sex	1.33 (1.13-1.56)	1.56 (1.16-2.08)	26.7 (9.8, 43.6)
2. Model 1 plus marital status and employment	1.26 (1.07-1.49)	1.48 (1.10-1.99)	24.3 (7.2, 41.4)
3. Model 2 plus delusions / hallucinations, and depression	1.24 (1.05-1.45)	1.48 (1.10-1.99)	24.4 (7.5, 41.4)

*Logistic regression; **Linear regression

Age x negative symptoms (binary variable) interaction term p>0.05 for all models

5.2 Supplementary methods

5.21 SQL data extraction

Three separate tables were created for each of the samples analysed in the study (described fully in section 5.1):

(i) Sample A: patients receiving care from SLaM in 2011 with a diagnosis of schizophrenia.

(ii) Sample B: a subset of Sample A who were in hospital during 2011 and discharged during 2011.

(iii) Sample C: a subset of Sample A who were admitted to hospital during 2011.

The SQL query below describes data extraction of covariates in sample A (from which samples B and C were also derived).

```
select ds1.BrcId
,floor((datediff (day,patient.cleaneddateofbirth , cast ('01/01/2011' as
datetime))/365)) age
,Gender_ID sex
,Marital_Status_ID
,case when emp1.Assessed_Date is not null then case when emp.Assessed_Date
is not null then case when
      abs(datediff(d,emp.Assessed_Date,'1-Jan-2011'))>=abs(datediff(d,'1-
Jan-2011',emp1.Assessed_Date))
      then emp1.PSA_Employment_Status_ID
      else emp.PSA_Employment_Status_ID end
      else emp1.PSA_Employment_Status_ID end
      else case when emp.PSA_Employment_Status_ID is not null then
      emp.PSA_Employment_Status_ID
      else null end end as Employment
,case when emp1.Assessed_Date is not null then case when emp.Assessed_Date
is not null then case when
      abs(datediff(d,emp.Assessed_Date,'1-Jan-2011'))>=abs(datediff(d,'1-
Jan-2011',emp1.Assessed_Date))
      then emp1.Assessed_Date
      else emp.Assessed_Date end
      else emp1.Assessed_Date end
      else case when emp.Assessed_Date is not null then
      emp.Assessed_Date
      else null end end as Employment_status_date
```

```
,case when honos1.rating_date is not null then case when honos.rating_date
is not null then case when
      abs(datediff(d,honos.rating_date,'1-Jan-2011'))>=abs(datediff(d,'1-
Jan-2011',honos1.rating_date))
      then honos1.HONOSTYPE
      else honos.HONOSTYPE end
      else honos1.HONOSTYPE end
      else case when honos.rating_date is not null then
      honos.HONOSTYPE
      else null end end as honostype
,case when honos1.rating_date is not null then case when honos.rating_date
is not null then case when
      abs(datediff(d,honos.rating_date,'1-Jan-2011'))>=abs(datediff(d,'1-
Jan-2011',honos1.rating_date))
      then honos1.Rating_Date
      else honos.Rating_Date end
      else honos1.Rating_Date end
      else case when honos.rating_date is not null then
      honos.Rating_Date
      else null end end as [Rating_Date]
,case when honos1.rating_date is not null then case when honos.rating_date
is not null then case when
      abs(datediff(d,honos.rating_date,'1-Jan-2011'))>=abs(datediff(d,'1-
Jan-2011',honos1.rating_date))
      then honos1.Hallucinations_Score_ID
      else honos.Hallucinations_Score_ID end
      else honos1.Hallucinations_Score_ID end
      else case when honos.rating_date is not null then
      honos.Hallucinations_Score_ID
      else null end end as Hallucinations_Score_ID
,case when honos1.rating_date is not null then case when honos.rating_date
is not null then case when
      abs(datediff(d,honos.rating_date,'1-Jan-2011'))>=abs(datediff(d,'1-
Jan-2011',honos1.rating_date))
      then honos1.Depressed_Mood_Score_ID
      else honos.Depressed_Mood_Score_ID end
      else honos1.Depressed_Mood_Score_ID end
      else case when honos.rating_date is not null then
      honos.Depressed_Mood_Score_ID
      else null end end  as Depressed_Mood_Score_ID
,case when honos1.rating_date is not null then case when honos.rating_date
is not null then case when
      abs(datediff(d,honos.rating_date,'1-Jan-2011'))>=abs(datediff(d,'1-
Jan-2011',honos1.rating_date))
      then honos1.Relationship_Problems_Score_ID
      else honos.Relationship_Problems_Score_ID end
      else honos1.Relationship_Problems_Score_ID end
      else case when honos.rating_date is not null then
      honos.Relationship_Problems_Score_ID
```

```
      else null end end as Relationship_Problems_Score_ID
,case when honos1.rating_date is not null then case when honos.rating_date
is not null then case when
      abs(datediff(d,honos.rating_date,'1-Jan-2011'))>=abs(datediff(d,'1-
Jan-2011',honos1.rating_date))
      then honos1.Daily_Living_Problems_Score_ID
      else honos.Daily_Living_Problems_Score_ID end
      else honos1.Daily_Living_Problems_Score_ID end
      else case when honos.rating_date is not null then
      honos.Daily_Living_Problems_Score_ID
      else null end end  as Daily_Living_Problems_Score_ID
,Inpatientin2011 = case when wsperiod.brcid is not null then 'Yes' else
'No' end
,case when Dep.brcid is not null then 'Yes' else 'No' end Depression
from
      (select
      distinct brcid
      from sqlcris.dbo.Referral r
      where Accepted_Date between '01-jan-1999' and '31-dec-2011' and
      (Discharge_Date >= '01-jan-2011' or Discharge_Date = '01-jan-1900')
and
      Referral_Status_ID not like 'rejected' and
      exists(select * from tbl_rs_f20count1 where brcid=r.brcid)
      ) ds1
join SQLCRIS.dbo.epr_form patient on ds1.BrcId=patient.brcid
left join [SQLCrisImport].[dbo].[vw_honos_source] honos with (nolock) on
      honos.cn_doc_id=(select top 1 cn_doc_id from
[SQLCrisImport].[dbo].[vw_honos_source] with (nolock)
      where  brcid=ds1.brcid and rating_date<='1-Jan-2011' and
Invalid_Flag_ID<>'Yes' order by rating_date desc,CN_Doc_ID desc)
left join [SQLCrisImport].[dbo].[vw_honos_source] honos1 with (nolock) on
      honos1.cn_doc_id=(select top 1 cn_doc_id from
[SQLCrisImport].[dbo].[vw_honos_source] with (nolock)
      where  brcid=ds1.brcid and rating_date>'1-Jan-2011' and
Invalid_Flag_ID<>'Yes' order by rating_date asc,CN_Doc_ID desc)
left join [SQLCris].[dbo].[Summary_of_Need] emp with (nolock) on
      emp.cn_doc_id=(select top 1 cn_doc_id from
[SQLCris].[dbo].[Summary_of_Need] with (nolock)
      where brcid=ds1.brcid and Assessed_Date<='1-Jan-2011' and
Invalid_Flag_ID<>'Yes' order by Assessed_Date desc,CN_Doc_ID desc)
left join [SQLCris].[dbo].[Summary_of_Need]  emp1 with (nolock) on
      emp1.cn_doc_id=(select top 1 cn_doc_id from
[SQLCris].[dbo].[Summary_of_Need] with (nolock)
      where brcid=ds1.brcid and Assessed_Date>'1-Jan-2011' and
Invalid_Flag_ID<>'Yes' order by Assessed_Date asc,CN_Doc_ID desc)
left join (select distinct BrcId from  sqlcris.dbo.ward_stay a
      where ((a.Current_Ward_Stay_Status_ID like 'closed' or
      a.Current_Ward_Stay_Status_ID like '%occupied%') and
location_name<>'Test Ward' and
```

```
      Actual_Start_Date > '01-jan-1900' and
      ('1-jan-2011' between actual_start_date and case when
actual_end_date='1-jan-1900' then GETDATE() else actual_end_date end))
      ) ws on ws.BrcId =ds1.BrcId
left join (select  distinct brcid from  sqlcris.dbo.ward_stay a
      where ((a.Current_Ward_Stay_Status_ID like 'closed' or
      a.Current_Ward_Stay_Status_ID like '%occupied%') and
location_name<>'Test Ward' and
      ((Actual_Start_Date > '01-jan-1900' and  Actual_Start_Date <= '31-
dec-2011') and
      (Actual_End_Date >= '01-jan-2011' or Actual_End_Date = '01-jan-
1900')))
      ) wsperiod on ds1.BrcId =wsperiod.brcid
left join vw_rs_depression dep on dep.brcid=ds1.brcid
```

This query extracted data for patients receiving mental healthcare from SLaM between 1st January 2011 and 31st December 2011 from the "tbl_rs_f20count1" table, which contains a cohort of all patients in the BRC Case Register with a diagnosis of schizophrenia (defined using the methods described in section 2.51). All covariates were extracted closest to 1st January 2011. Age, gender and marital status were extracted using a table join to the "SQLCRIS.dbo.epr_form" table. Employment status was extracted using a CASE WHEN statements with a table join to the "[SQLCris].[dbo].[Summary_of_Need]" table. This join selects the closest record occurring prior to 1st January 2011. HoNOS scores for hallucinations and delusions ("honos.Hallucinations_Score_ID"), depressed mood ("honos.Depressed_Mood_Score_ID"), problems with relationships ("honos.Relationship_Problems_Score_ID") and ADL impairment ("honos.Daily_Living_Problems_Score_ID") were extracted using CASE WHEN statements with a table join to the "[SQLCrisImport].[dbo].[vw_honos_source]" table. A CASE WHEN statement with a table join to the "sqlcris.dbo.ward_stay" table was used to determine whether patients received inpatient care during 2011. As well as using HoNOS data to determine the presence of depressed mood, a table join with a CASE WHEN statement was used upon the "vw_rs_depression" view. This view contains all patients in the BRC Case Register with a diagnosis of depression (defined using the methods described in section 2.51).

5.22 SQL support queries

In order to extract the presence of negative symptoms from each sample analysed in the study (A, B and C), the following SQL query was performed.

```
USE [SQLCrisImport]
SELECT a.*
      ,
     abstract_think_concrete,
     abstract_think_concrete_high_prob
    ,Affect_abnormal
    ,Affect_blunted
    ,Affect_flat
    ,Affect_reactive
    ,Apathy_apathetic
    ,EW_withdrawn
    ,Eye_contact_good
    ,Eye_contact_intermediate
    ,Eye_contact_poor
    ,POS_abnormal
    ,POS_normal
    ,POS_poverty
    ,Rapport_good
    ,Rapport_poor
    ,SW_socially_withdrawn
    ,motivation
    ,mutism
    ,neg_symptoms
    ,neg_symptoms_high_prob

  FROM [SQLCrisImport].[dbo].[vw_rs_ns_sample1_new] a
left join
(SELECT
brcid,
COUNT(*) abstract_think_concrete
  FROM
  [GateDB_Cris].[dbo].[gate_abstract_thinking_current]
  where convert (datetime, document_date, 103) < '01-jan-2012'
  group by BrcId) atc on a.brcid=atc.brcid

left join
(SELECT
brcid,
```

```
COUNT(*) abstract_think_concrete_high_prob
  FROM
  [GateDB_Cris].[dbo].[gate_abstract_thinking_current]
  where convert (datetime, document_date, 103) < '01-jan-2012'
  and
  convert (real, [prob]) >= '0.7'
  group by BrcId) atc1 on a.brcid=atc1.brcid

left join
(SELECT
brcid,
COUNT(*) Affect_abnormal
  FROM GateDB_Cris.dbo.gate_affect_current
  where affect = 'abnormal'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) aa on a.brcid=aa.brcid

left join
(SELECT
brcid,
COUNT(*) Affect_blunted
  FROM GateDB_Cris.dbo.gate_affect_current
  where affect = 'blunted'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) ab on a.brcid=ab.brcid

left join
(SELECT
brcid,
COUNT(*) Affect_flat
  FROM GateDB_Cris.dbo.gate_affect_current
  where affect = 'flat'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) af on a.brcid=af.brcid

left join
(SELECT
brcid,
COUNT(*) Affect_reactive
  FROM GateDB_Cris.dbo.gate_affect_current
  where affect = 'reactive'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) ar on a.brcid=ar.brcid

left join
(SELECT
brcid,
COUNT(*) Apathy_apathetic
  FROM GateDB_Cris.dbo.gate_apathy_current
```

```
  where convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) aap on a.brcid=aap.brcid

left join
(SELECT
brcid,
COUNT(*) EW_withdrawn
  FROM GateDB_Cris.dbo.gate_emotional_withdrawal_current
  where convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) ew on a.brcid=ew.brcid

left join
(SELECT
brcid,
COUNT(*) Eye_contact_good
  FROM GateDB_Cris.dbo.gate_eye_contact_current
  where eye_contact = 'good'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) ecg on a.brcid=ecg.brcid

left join
(SELECT
brcid,
COUNT(*) Eye_contact_intermediate
  FROM GateDB_Cris.dbo.gate_eye_contact_current
  where eye_contact = 'intermediate'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) eci on a.brcid=eci.brcid

left join
(SELECT
brcid,
COUNT(*) Eye_contact_poor
  FROM GateDB_Cris.dbo.gate_eye_contact_current
  where eye_contact = 'poor'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) ecp on a.brcid=ecp.brcid

left join
(SELECT
brcid,
COUNT(*) POS_abnormal
  FROM GateDB_Cris.dbo.gate_poverty_of_speech_current
  where poverty_of_speech = 'abnormal'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) posa  on a.brcid=posa.brcid

left join
(SELECT
```

```
brcid,
COUNT(*) POS_normal
  FROM GateDB_Cris.dbo.gate_poverty_of_speech_current
  where poverty_of_speech = 'normal'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) pos on a.brcid=pos.brcid

left join
(SELECT
brcid,
COUNT(*) POS_poverty
  FROM GateDB_Cris.dbo.gate_poverty_of_speech_current
  where poverty_of_speech = 'poverty'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) posp on a.brcid=posp.brcid

left join
(SELECT
brcid,
COUNT(*) Rapport_good
  FROM GateDB_Cris.dbo.gate_rapport_current
  where rapport = 'good'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) rg on a.brcid=rg.brcid

left join
(SELECT
brcid,
COUNT(*) Rapport_poor
  FROM GateDB_Cris.dbo.gate_rapport_current
  where rapport = 'poor'
  and convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) rp on  a.brcid=rp.brcid

left join
(SELECT
brcid,
COUNT(*) SW_socially_withdrawn
  FROM GateDB_Cris.dbo.gate_social_withdrawal_current
  where convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) ssw on a.brcid=ssw.brcid

left join
(SELECT
brcid,
COUNT(*) motivation
  FROM GateDB_Cris.dbo.gate_motivation_current
  where convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) mtv on a.brcid=mtv.brcid
```

```
left join
(SELECT
brcid,
COUNT(*) mutism
  FROM GateDB_Cris.dbo.gate_mutism_current
  where convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) mts on a.brcid=mts.brcid

left join
(SELECT
brcid,
COUNT(*) neg_symptoms
  FROM GateDB_Cris.dbo.gate_negative_symptoms_current
  where convert (datetime, document_date, 103) < '01-jan-2012'
  group by brcid) ns on a.brcid=ns.brcid

left join
(SELECT
brcid,
COUNT(*) neg_symptoms_high_prob
  FROM GateDB_Cris.dbo.gate_negative_symptoms_current
  where convert (datetime, document_date, 103) < '01-jan-2012'
  and convert (real, [prob]) >= '0.9'
  group by brcid) nsh on a.brcid=nsh.brcid
```

The SQL query extracted negative symptoms data by means of left outer joins to tables containing NLP data for each symptom upon the "[SQLCrisImport].[dbo].[vw_rs_ns_sample1_new]" view, which contains all patients included in Sample A. The same script was used to join negative symptoms data to Sample B and Sample C. Each table join returned a count of the number of mentions of a particular symptom recorded prior to 1st January 2012. These variables were subsequently recoded using the method described in section 5.24 to generate the binary exposures in the ten point CRIS-NSS scale which was analysed in the study. The method for populating each source table for negative symptoms in the GateDB_Cris database is described in section 5.23.

5.23 Negative symptoms NLP development

The negative symptoms analysed in this study were ascertained using ten NLP applications developed with SVM. A further rules-based algorithm was applied to a subset of the SVM NLP applications to further enhance their accuracy. Full details of the NLP development are provided in a previously published article.[121] In summary, eleven applications were developed based on a range of key words relevant to negative symptoms:

(i) Concrete thinking: concrete thinking present or absent (2 classes, 118 sentences)

(ii) Affect: reactive, normal, abnormal, blunted, flat (5 classes, 103 sentences)

(iii) Apathy: apathy present or absent (2 classes, 145 sentences)

(iv) Emotional withdrawal: withdrawn, detached, no emotional withdrawal (3 classes, 118 sentences)

(v) Eye contact: good, intermediate, poor, unknown (4 classes, 35 sentences)

(vi) Motivation: poor, unknown (2 classes, 259 sentences)

(vii) Mutism: mutism present or absent (2 classes, 234 sentences)

(viii) Negative symptoms: negative symptoms present or absent (2 classes, 185 sentences)

(ix) Poverty of speech: normal, abnormal, poverty, unknown(4 classes, 263 sentences)

(x) Rapport: good, poor, unknown (3 classes, 139 sentences)

(xi) Social withdrawal: social withdrawal present or absent (2 classes, 166 sentences)

The number of classes and number of sentences used in the seed training dataset are given in parentheses. For the purposes of this study, only applications extracting a specific negative symptom were used to analyse the data (i.e. application (viii) looking at overall “negative symptoms” was not used). This was to allow for analysis of individual symptoms without the risk of collinearity with an NLP application extracting “negative symptoms” as a whole. The NLP applications were developed using GATE-ML software with a “bag-of-words” SVM approach and validated using 5 fold cross validation. A gold standard reference dataset was not used for validation purposes and the precision and recall statistics quoted in section 5.1 represent the best overall results (measured as the F1, the harmonic mean of precision and recall) generated through 5 fold cross validation.

Following the first round of SVM NLP development, one round of active learning was applied to the concrete thinking (99 sentences), affect (200 sentences), emotional withdrawal (100 sentences), poverty of speech (62 sentences) and rapport (37 sentences) applications. All training data included in the development of the ten applications included in the study (2263 sentences) were annotated by two psychiatrists (Rashmi Patel and Rob Stewart) resulting in percentage inter-annotator agreement of 93% and Cohen's kappa of 0.85. An SVM margin filter of 0.4 was applied to all applications to ensure all applications reached a minimum precision threshold of 80%. In order to further optimise precision statistics, rules-based queries were applied using the Java Annotation Patterns Engine (JAPE) to the following applications: concrete thinking, affect, emotional withdrawal, eye contact, mutism, poverty of speech and rapport. The best results of the ten applications are presented in Table 1 in section 5.1.

After optimising the NLP applications using active learning and superimposition of rules, the applications were run over correspondence documents in the BRC Case Register. SQL queries (described in section 5.21) were used to join the output data from the NLP applications onto data from a cohort of patients with schizophrenia to obtain the data presented in the study. Data on the following classes were obtained:

(i) Concrete thinking: concrete thinking present

(ii) Affect: blunted or flat

(iii) Apathy: apathy present

(iv) Emotional withdrawal: withdrawn or detached

(v) Eye contact: intermediate or poor

(vi) Motivation: poor

(vii) Mutism: mutism present

(viii) Poverty of speech: poverty

(ix) Rapport: poor

(x) Social withdrawal: social withdrawal present

5.24 Statistical analysis

Three sets of analyses were performed: Sample A, Sample B and Sample C (described further in the main manuscript in section 5.1). For each analysis, a common set of STATA commands was performed in order to prepare data for subsequent analysis. Unlike the studies presented in chapters 4 and 5, conversion of string variables to numerical variables was performed in STATA. This was followed by recoding of numerical variables into the categories analysed in the study. The STATA commands which were used are reproduced below:

```
*Recoding GATE symptoms*

gen new_concrete = abstract_think_concrete
replace new_concrete="0" if abstract_think_concrete=="NULL"
destring new_concrete, replace

gen new_aff_abn = affect_abnormal
replace new_aff_abn = "0" if affect_abnormal=="NULL"
destring new_aff_abn, replace

gen new_aff_blunt = affect_blunted
replace new_aff_blunt = "0" if affect_blunted=="NULL"
destring new_aff_blunt, replace

gen new_aff_flat = affect_flat
replace new_aff_flat = "0" if affect_flat=="NULL"
destring new_aff_flat, replace

gen new_aff_react = affect_reactive
replace new_aff_react = "0" if affect_reactive=="NULL"
destring new_aff_react, replace

gen new_apathy = apathy_apathetic
replace new_apathy = "0" if apathy_apathetic=="NULL"
destring new_apathy, replace

gen new_ew = ew_withdrawn
replace new_ew = "0" if ew_withdrawn=="NULL"
destring new_ew, replace
```

```
gen new_ec_good = eye_contact_good
replace new_ec_good = "0" if eye_contact_good=="NULL"
destring new_ec_good, replace

gen new_ec_int = eye_contact_intermediate
replace new_ec_int = "0" if eye_contact_intermediate=="NULL"
destring new_ec_int, replace

gen new_ec_poor = eye_contact_poor
replace new_ec_poor = "0" if eye_contact_poor=="NULL"
destring new_ec_poor, replace

gen new_speech_abn = pos_abnormal
replace new_speech_abn = "0" if pos_abnormal=="NULL"
destring new_speech_abn, replace

gen new_speech_norm = pos_normal
replace new_speech_norm = "0" if pos_normal=="NULL"
destring new_speech_norm, replace

gen new_speech_poor = pos_poverty
replace new_speech_poor = "0" if pos_poverty=="NULL"
destring new_speech_poor, replace

gen new_rapp_good = rapport_good
replace new_rapp_good = "0" if rapport_good=="NULL"
destring new_rapp_good, replace

gen new_rapp_poor = rapport_poor
replace new_rapp_poor = "0" if rapport_poor=="NULL"
destring new_rapp_poor, replace

gen new_sw = sw_socially_withdrawn
replace new_sw = "0" if sw_socially_withdrawn=="NULL"
destring new_sw, replace

gen new_amotiv = motivation
replace new_amotiv = "0" if motivation=="NULL"
destring new_amotiv, replace

gen new_mute = mutism
replace new_mute = "0" if mutism=="NULL"
destring new_mute, replace

gen new_neg1 = neg_symptoms
replace new_neg1 = "0" if neg_symptoms=="NULL"
destring new_neg1, replace

gen new_neg2 = neg_symptoms_high_prob
```

```
replace new_neg2 = "0" if neg_symptoms_high_prob=="NULL"
destring new_neg2, replace

*CRIS-NSS scales*

gen GATEn1a = new_aff_blunt
recode GATEn1a 1/max=1

gen GATEn2a = new_aff_flat
recode GATEn2a 1/max=1

gen GATEn1a2a = (new_aff_blunt + new_aff_flat)
recode GATEn1a2a 1/max=1

gen GATEn3a = new_ew
recode GATEn3a 1/max=1

gen GATEn4a = new_ec_int
recode GATEn4a 1/max=1

gen GATEn5a = new_ec_poor
recode GATEn5a 1/max=1

gen GATEn4a5a = (new_ec_int + new_ec_poor)
recode GATEn4a5a 1/max=1

gen GATEn6a = new_rapp_poor
recode GATEn6a 1/max=1

gen GATEn7a = new_apathy
recode GATEn7a 1/max=1

gen GATEn8a = new_sw
recode GATEn8a 1/max=1

gen GATEn9a = new_amotiv
recode GATEn9a 1/max=1

gen GATEn10a = new_concrete
recode GATEn10a 1/max=1

gen GATEn11a = new_speech_poor
recode GATEn11a 1/max=1

gen GATEn12a = new_mute
recode GATEn12a 1/max=1

alpha GATEn9a GATEn1a2a GATEn4a5a GATEn3a GATEn6a GATEn8a GATEn11a GATEn12a
GATEn7a GATEn10a, generate(GATEntotc) item
```

```
gen GATE10point = (GATEntotc)*10

*Recodings - generating numeric variables*

gen Sex = sex
replace Sex = "0" if sex=="Female"
replace Sex = "1" if sex=="Male"
replace Sex = "9" if sex=="Not Known"
replace Sex = "9" if sex=="Not Specified"
destring Sex, replace
recode Sex 9=.

gen Marit = marital_status_id
replace Marit = "1" if marital_status_id=="Cohabiting"
replace Marit = "2" if marital_status_id=="Divorced"
replace Marit = "3" if marital_status_id=="Divorced/Civil Partnership
Dissolved"
replace Marit = "4" if marital_status_id=="Married"
replace Marit = "5" if marital_status_id=="Married/Civil Partner"
replace Marit = "6" if marital_status_id=="Not Disclosed"
replace Marit = "7" if marital_status_id=="Not Known"
replace Marit = "8" if marital_status_id=="Separated"
replace Marit = "9" if marital_status_id=="Single"
replace Marit = "10" if marital_status_id=="Widowed"
replace Marit = "11" if marital_status_id=="Widowed/Surviving Civil
Partner"
destring Marit, replace

gen Employ = employment
replace Employ = "1" if employment=="Employed"
replace Employ = "2" if employment=="NULL"
replace Employ = "3" if employment=="Not applicable"
replace Employ = "4" if employment=="Not disclosed"
replace Employ = "5" if employment=="Not known"
replace Employ = "6" if employment=="Other employment status such as in
education or training"
replace Employ = "7" if employment=="Retired"
replace Employ = "8" if employment=="Unemployed"
replace Employ = "9" if employment=="xNx"
replace Employ = "10" if employment=="Long-term sick or disabled"
replace Employ = "11" if employment=="Students that are not actively
seeking work"
replace Employ = "12" if employment=="Unpaid voluntary work and not
actively seeking work"
destring Employ, replace

gen dep = depression
replace dep = "0" if depression=="No"
```

```
replace dep = "1" if depression=="Yes"
destring dep, replace

*Recodings - creating grouped variables*

gen agegp=age
recode agegp min/15=. 16/19=1 20/29=2 30/39=3 40/49=4 50/59=5 60/69=6
70/max=7
*i.e. number equals decade, below 16 dropped

gen agegp2=age
recode agegp2 min/15=. 16/39=0 40/max=1
*i.e. number equals decade, below 16 dropped

gen Sex3 = sex
replace Sex3 = "0" if sex=="Female"
replace Sex3 = "1" if sex=="Male"
replace Sex3 = "2" if sex=="Not Known"
replace Sex3 = "2" if sex=="Not Specified"
destring Sex3, replace
*0 Female
*1 Male
*2 Not recorded

gen Marit3 = Marit
recode Marit3 1=1 2=2 3=2 4=1 5=1 6=. 7=. 8=2 9=3 10=2 11=2
*1 Cohabiting/married
*2 Divorced/separated
*3 Single

gen Marit4 = Marit
recode Marit4 1=1 2=2 3=2 4=1 5=1 6=. 7=. 8=2 9=3 10=4 11=4
*1 Cohabiting/married
*2 Divorced/separated
*3 Single
*4 Widowed

gen Marit5 = Marit
recode Marit5 1=1 2=2 3=2 4=1 5=1 6=5 7=5 8=2 9=3 10=4 11=4
*1 Cohabiting/married
*2 Divorced/separated
*3 Single
*4 Widowed
*5 Not recorded

gen Employ4 = Employ
recode Employ4 1=1 2=. 3=. 4=. 5=. 6=2 7=3 8=0 9=. 10=0 11=2 12=0
*0 Unemployed
*1 Employed
```

```
*2 In education
*3 Retired

gen Employ5 = Employ
recode Employ5 1=1 2=4 3=4 4=4 5=4 6=2 7=3 8=0 9=4 10=0 11=2 12=0
*0 Unemployed
*1 Employed
*2 In education
*3 Retired
*4 Not recorded

*HoNOS recodings*

gen Honos_psychosis = hallucinations_score_id
replace Honos_psychosis = "9" if hallucinations_score_id=="NULL"
replace Honos_psychosis = "9" if hallucinations_score_id=="Missing"
destring Honos_psychosis, replace
recode Honos_psychosis 9=.
recode Honos_psychosis 1=0 2/4=1

gen Honos_psychosis3 = Honos_psychosis
recode Honos_psychosis3 0=0 1=1 .=2
*0 No
*1 Yes
*2 Not recorded

gen Honos_dep = depressed_mood_score_id
replace Honos_dep = "9" if depressed_mood_score_id=="NULL"
replace Honos_dep = "9" if depressed_mood_score_id=="Missing"
destring Honos_dep, replace
recode Honos_dep 9=.
recode Honos_dep 1=0 2/4=1

gen Honos_dep3 = Honos_dep
recode Honos_dep3 0=0 1=1 .=2
*0 No
*1 Yes
*2 Not recorded

gen Honos_soc = relationship_problems_score_id
replace Honos_soc = "9" if relationship_problems_score_id=="NULL"
replace Honos_soc = "9" if relationship_problems_score_id=="Missing"
destring Honos_soc, replace
recode Honos_soc 9=.
recode Honos_soc 1=0 2/4=1

gen Honos_soc3 = Honos_soc
recode Honos_soc3 0=0 1=1 .=2
*0 No
```

```
*1 Yes
*2 Not recorded

gen Honos_adl = daily_living_problems_score_id
replace Honos_adl = "9" if daily_living_problems_score_id=="NULL"
replace Honos_adl = "9" if daily_living_problems_score_id=="Missing"
destring Honos_adl, replace
recode Honos_adl 9=.
recode Honos_adl 1=0 2/4=1

gen Honos_adl3 = Honos_adl
recode Honos_adl3 0=0 1=1 .=2
*0 No
*1 Yes
*2 Not recorded

gen dep_any = dep + Honos_dep
recode dep_any 2=1

gen dep_any3 = dep_any
recode dep_any3 0=0 1=1 .=2
*0 No
*1 Yes
*2 Not recorded

*Negative symptom definition*

gen GATE10point2plus=GATE10point
recode GATE10point2plus 1=0 2/max=1

*Inclusion criteria for main analyses*

gen xx=1
recode xx 1=0 if age<16
```

The first set of commands (*Recoding GATE symptoms*) converted the negative symptom NLP data from SQL into numerical variables for each construct. The next set of commands (*CRIS-NSS scales*) recoded these NLP variables into the categories to be used in the CRIS-NSS scale of ten negative symptoms described in the main manuscript (section 5.1). The “alpha” command was used to estimate the Cronbach alpha value for the composite scale of ten negative symptoms and derive a new variable (“GATEntotc”) containing these data. This variable consisted of a scale from 0 to 1 in 0.1 increments for each subdomain of the scale. A new variable was generated (“GATE10point”) to transform this into a scale from 0 to 10. The next set of commands (*Recodings - generating numeric variables*) converted the text of covariates into numerical variables. This is analogous to the CASE WHEN statement method employed in SQL to recode data. The subsequent set of commands (*Recodings - creating grouped variables* and *HoNOS recodings*) recoded these covariates into categories employed in the analyses presented in the main manuscript (section 5.1). A new binary variable was generated (“GATE10point2plus”) to set a threshold of 2 or more negative symptoms in the CRIS-NSS scale to be used as an exposure for binary logistic and linear regression analyses. Finally, a selection variable “xx” was generated to create a filter to include only patients aged 16 years and older in the subsequent analyses.

The following analysis was performed on sample A to generate the results presented in the respective tables of the main manuscript and supplementary material (section 5.1). Details of the statistical methods employed to generate the data in the tables is provided in the main manuscript (section 5.1).

```
*Table 1 - negative symptoms prevalences*

tab GATEn1a2a if xx==1
tab GATEn3a if xx==1
tab GATEn4a5a if xx==1
tab GATEn6a if xx==1
tab GATEn7a if xx==1
tab GATEn8a if xx==1
tab GATEn9a if xx==1
tab GATEn10a if xx==1
tab GATEn11a if xx==1
tab GATEn12a if xx==1
tab GATE10point if xx==1

*Table 2 - 10 point including missing*

tab agegp GATE10point2plus if xx==1, row
tab Sex3 GATE10point2plus if xx==1, row
tab Marit5 GATE10point2plus if xx==1, row
tab Employ5 GATE10point2plus if xx==1, row
tab Honos_adl3 GATE10point2plus if xx==1, row
```

```
tab Honos_soc3 GATE10point2plus if xx==1, row
tab Honos_psychosis3 GATE10point2plus if xx==1, row
tab dep_any3 GATE10point2plus if xx==1, row

logistic GATE10point2plus ib2.agegp if xx==1
logistic GATE10point2plus ib1.Sex3 if xx==1
logistic GATE10point2plus ib3.Marit5 if xx==1
logistic GATE10point2plus i.Employ5 if xx==1
logistic GATE10point2plus i.Honos_adl3 if xx==1
logistic GATE10point2plus i.Honos_soc3 if xx==1
logistic GATE10point2plus i.Honos_psychosis3 if xx==1
logistic GATE10point2plus i.dep_any3 if xx==1

*Table 2 - 10 point multivariable binary logistic regression including
missing*
*N.B. Sex variable not including missing as only 2 cases*

logistic GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
i.Honos_adl3 i.Honos_soc3 i.Honos_psychosis3 i.dep_any3 if xx==1

*Table 3 - adjusted inpatient admission regression including missing*
*N.B. Sex variable not including missing as only 2 cases*
logistic ip2011any GATE10point2plus if xx==1
logistic ip2011any GATE10point2plus ib2.agegp ib1.Sex if xx==1
logistic ip2011any GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
if xx==1
logistic ip2011any GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1

logistic ip2011any GATE10point if xx==1
logistic ip2011any GATE10point ib2.agegp ib1.Sex if xx==1
logistic ip2011any GATE10point ib2.agegp ib1.Sex ib3.Marit5 i.Employ5 if
xx==1
logistic ip2011any GATE10point ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1

*eTable 1 - 10 point ordinal logistic regression including missing*
*N.B. Sex variable not including missing as only 2 cases*
ologit GATE10point ib2.agegp if xx==1, or
ologit GATE10point ib1.Sex if xx==1, or
ologit GATE10point ib3.Marit5 if xx==1, or
ologit GATE10point i.Employ5 if xx==1, or
ologit GATE10point i.Honos_adl if xx==1, or
ologit GATE10point i.Honos_soc if xx==1, or
ologit GATE10point i.Honos_psychosis if xx==1, or
ologit GATE10point i.dep_any3 if xx==1, or

*eTable 1 - 10 point multivariable ordinal logistic regression including
missing*
```

5. Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method

```
*N.B. Sex variable not including missing as only 2 cases*
ologit GATE10point ib2.agegp ib1.Sex ib3.Marit5 i.Employ5 i.Honos_adl
i.Honos_soc i.Honos_psychosis i.dep_any3 if xx==1, or

*eTable 2 - 10 point with missing data dropped*

tab agegp GATE10point2plus if xx==1, row
tab Sex GATE10point2plus if xx==1, row
tab Marit4 GATE10point2plus if xx==1, row
tab Employ4 GATE10point2plus if xx==1, row
tab Honos_adl GATE10point2plus if xx==1, row
tab Honos_soc GATE10point2plus if xx==1, row
tab Honos_psychosis GATE10point2plus if xx==1, row
tab dep_any GATE10point2plus if xx==1, row

logistic GATE10point2plus ib2.agegp if xx==1
logistic GATE10point2plus ib1.Sex if xx==1
logistic GATE10point2plus ib3.Marit4 if xx==1
logistic GATE10point2plus i.Employ4 if xx==1
logistic GATE10point2plus i.Honos_adl if xx==1
logistic GATE10point2plus i.Honos_soc if xx==1
logistic GATE10point2plus i.Honos_psychosis if xx==1
logistic GATE10point2plus i.dep_any if xx==1

*eTable 2 - 10 point multivariable binary logistic regression with missing
data dropped*
logistic GATE10point2plus ib2.agegp ib1.Sex ib3.Marit4 i.Employ4
i.Honos_adl i.Honos_soc i.Honos_psychosis i.dep_any if xx==1

*eTable 3 - 10 point inpatient admission during 2011 descriptive
statistics*
tab GATE10point ip2011any if xx==1, row

*eTable 6 - Age interaction analysis including missing*
*N.B. Sex variable not including missing as only 2 cases*
logistic ip2011any GATE10point2plus if xx==1&age<=39
logistic ip2011any GATE10point2plus if xx==1&age>=40
logistic ip2011any GATE10point2plus GATE10point2plus#agegp2 if xx==1

logistic ip2011any GATE10point2plus ib2.agegp ib1.Sex if xx==1&age<=39
logistic ip2011any GATE10point2plus ib2.agegp ib1.Sex if xx==1&age>=40
logistic ip2011any GATE10point2plus ib2.agegp GATE10point2plus#agegp2
ib1.Sex if xx==1

logistic ip2011any GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
if xx==1&age<=39
logistic ip2011any GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
if xx==1&age>=40
```

5. Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method

```
logistic ip2011any GATE10point2plus ib2.agegp GATE10point2plus#agegp2
ib1.Sex ib3.Marit5 i.Employ5 if xx==1

logistic ip2011any GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1&age<=39
logistic ip2011any GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1&age>=40
logistic ip2011any GATE10point2plus ib2.agegp GATE10point2plus#agegp2
ib1.Sex ib3.Marit5 i.Employ5 Honos_psychosis3 dep_any3 if xx==1
```

These commands were applied to two variations of sample A. The first variation included negative symptoms recorded at any time point in the electronic health record. This was used to generate the data in Table 1, Table 2, eTable 1 and eTable 2. The second variation only included negative symptoms recorded prior to 1st January 2011 in order to ensure these occurred prior to the outcomes measured in Table 3, eTable 3 and eTable 6.

The following analysis was performed on sample B to generate the results presented in the respective tables of the main manuscript and supplementary material (section 5.1) Details of the statistical methods employed to generate the data in the tables is provided in the main manuscript (section 5.1).

```
**** Analyses ****

*Table 3 - adjusted readmission regression including missing*
*N.B. Sex variable not including missing as only 2 cases*
logistic Readm GATE10point2plus if xx==1
logistic Readm GATE10point2plus ib2.agegp ib1.Sex if xx==1
logistic Readm GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5 if
xx==1
logistic Readm GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1

logistic Readm GATE10point if xx==1
logistic Readm GATE10point ib2.agegp ib1.Sex if xx==1
logistic Readm GATE10point ib2.agegp ib1.Sex ib3.Marit5 i.Employ5 if xx==1
logistic Readm GATE10point ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
```

```
*Table 4 - unadjusted readmission regression including missing*
logistic Readm GATEn1a2a if xx==1
logistic Readm GATEn3a if xx==1
logistic Readm GATEn4a5a if xx==1
logistic Readm GATEn6a if xx==1
logistic Readm GATEn7a if xx==1
logistic Readm GATEn8a if xx==1
logistic Readm GATEn9a if xx==1
logistic Readm GATEn10a if xx==1
logistic Readm GATEn11a if xx==1
logistic Readm GATEn12a if xx==1

*Table 4 - adjusted readmission regression including missing*
*N.B. Sex variable not including missing as only 2 cases*
logistic Readm GATEn1a2a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
logistic Readm GATEn3a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
logistic Readm GATEn4a5a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
logistic Readm GATEn6a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
logistic Readm GATEn7a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
logistic Readm GATEn8a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
logistic Readm GATEn9a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
logistic Readm GATEn10a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
logistic Readm GATEn11a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
logistic Readm GATEn12a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1

*eTable 4 - 10 point readmission descriptive statistics*
tab GATE10point Readm if xx==1, row

*eTable 6 - Age interaction analysis including missing*
*N.B. Sex variable not including missing as only 2 cases*
logistic Readm GATE10point2plus if xx==1&age<=39
logistic Readm GATE10point2plus if xx==1&age>=40
logistic Readm GATE10point2plus GATE10point2plus#agegp2 if xx==1

logistic Readm GATE10point2plus ib2.agegp ib1.Sex if xx==1&age<=39
logistic Readm GATE10point2plus ib2.agegp ib1.Sex if xx==1&age>=40
logistic Readm GATE10point2plus ib2.agegp GATE10point2plus#agegp2 ib1.Sex
if xx==1
```

5. Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method

```
logistic Readm GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5 if
xx==1&age<=39
logistic Readm GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5 if
xx==1&age>=40
logistic Readm GATE10point2plus ib2.agegp GATE10point2plus#agegp2 ib1.Sex
ib3.Marit5 i.Employ5 if xx==1

logistic Readm GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1&age<=39
logistic Readm GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1&age>=40
logistic Readm GATE10point2plus ib2.agegp GATE10point2plus#agegp2 ib1.Sex
ib3.Marit5 i.Employ5 Honos_psychosis3 dep_any3 if xx==1
```

The following analysis was performed on sample C to generate the results presented in the respective tables of the main manuscript and supplementary material (section 5.1) Details of the statistical methods employed to generate the data in the tables is provided in the main manuscript (section 5.1).

```
*Table 3 - adjusted length of stay regression including missing*
*N.B. Sex variable not including missing as only 2 cases*
regress lengthofstay GATE10point2plus if xx==1
regress lengthofstay GATE10point2plus ib2.agegp ib1.Sex if xx==1
regress lengthofstay GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5
i.Employ5 if xx==1
regress lengthofstay GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5
i.Employ5 Honos_psychosis3 dep_any3 if xx==1

regress lengthofstay GATE10point if xx==1
regress lengthofstay GATE10point ib2.agegp ib1.Sex if xx==1
regress lengthofstay GATE10point ib2.agegp ib1.Sex ib3.Marit5 i.Employ5 if
xx==1
regress lengthofstay GATE10point ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1

*Table 4 - unadjusted length of stay regression including missing*
regress lengthofstay GATEn1a2a if xx==1
regress lengthofstay GATEn3a if xx==1
regress lengthofstay GATEn4a5a if xx==1
regress lengthofstay GATEn6a if xx==1
```

```
regress lengthofstay GATEn7a if xx==1
regress lengthofstay GATEn8a if xx==1
regress lengthofstay GATEn9a if xx==1
regress lengthofstay GATEn10a if xx==1
regress lengthofstay GATEn11a if xx==1
regress lengthofstay GATEn12a if xx==1

*Table 4 - adjusted length of stay regression including missing*
*N.B. Sex variable not including missing as only 2 cases*
regress lengthofstay GATEn1a2a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
regress lengthofstay GATEn3a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
regress lengthofstay GATEn4a5a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
regress lengthofstay GATEn6a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
regress lengthofstay GATEn7a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
regress lengthofstay GATEn8a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
regress lengthofstay GATEn9a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
regress lengthofstay GATEn10a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
regress lengthofstay GATEn11a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1
regress lengthofstay GATEn12a ib2.agegp ib1.Sex ib3.Marit5 i.Employ5
Honos_psychosis3 dep_any3 if xx==1

*eTable 5 - 10 point length of stay descriptive statistics*
sort GATE10pointmax6
by GATE10pointmax6: summ lengthofstay, detail

*eTable 6 - Age interaction analysis including missing*
*N.B. Sex variable not including missing as only 2 cases*
regress lengthofstay GATE10point2plus if xx==1&age<=39
regress lengthofstay GATE10point2plus if xx==1&age>=40
regress lengthofstay GATE10point2plus GATE10point2plus#agegp2 if xx==1

regress lengthofstay GATE10point2plus ib2.agegp ib1.Sex if xx==1&age<=39
regress lengthofstay GATE10point2plus ib2.agegp ib1.Sex if xx==1&age>=40
regress lengthofstay GATE10point2plus ib2.agegp GATE10point2plus#agegp2
ib1.Sex if xx==1

regress lengthofstay GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5
i.Employ5 if xx==1&age<=39
regress lengthofstay GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5
i.Employ5 if xx==1&age>=40
```

5. Negative symptoms in schizophrenia: a study in a large clinical sample of patients using a novel automated method

```
regress lengthofstay GATE10point2plus ib2.agegp GATE10point2plus#agegp2
ib1.Sex ib3.Marit5 i.Employ5 if xx==1

regress lengthofstay GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5
i.Employ5 Honos_psychosis3 dep_any3 if xx==1&age<=39
regress lengthofstay GATE10point2plus ib2.agegp ib1.Sex ib3.Marit5
i.Employ5 Honos_psychosis3 dep_any3 if xx==1&age>=40
regress lengthofstay GATE10point2plus ib2.agegp GATE10point2plus#agegp2
ib1.Sex ib3.Marit5 i.Employ5 Honos_psychosis3 dep_any3 if xx==1
```

6. TextHunter - A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research

This chapter incorporates an article published in the Proceedings of the American Medical Informatics Association describing the NLP methods used in this thesis, including the cannabis NLP application described in chapter 4.

6.1 Proceedings of the American Medical Informatics Association article

Please see overleaf.

TextHunter – A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research

Richard G. Jackson MSc[1], Michael Ball MSc[1], Rashmi Patel BMBCh[1], Richard D. Hayes PhD[1], Richard J.B. Dobson PhD[1], Robert Stewart MD[1]

[1]King's College London (Institute of Psychiatry), London, UK

Abstract

Observational research using data from electronic health records (EHR) is a rapidly growing area, which promises both increased sample size and data richness - therefore unprecedented study power. However, in many medical domains, large amounts of potentially valuable data are contained within the free text clinical narrative. Manually reviewing free text to obtain desired information is an inefficient use of researcher time and skill. Previous work has demonstrated the feasibility of applying Natural Language Processing (NLP) to extract information. However, in real world research environments, the demand for NLP skills outweighs supply, creating a bottleneck in the secondary exploitation of the EHR. To address this, we present TextHunter, a tool for the creation of training data, construction of concept extraction machine learning models and their application to documents. Using confidence thresholds to ensure high precision (>90%), we achieved recall measurements as high as 99% in real world use cases.

Introduction

The increasing use of electronic health records (EHR) provides potentially transformative opportunities for clinical research in the breadth and depth of data contained within them. However, unstructured clinical notes are often the most valuable source of phenotypic/contextual information because of limitations in the scope and acceptability of structured fields. In response to this challenge, Natural Language Processing (NLP) has been employed to extract appropriate assertions in a structured format amenable to the needs of researchers[1]. While significant success has been achieved in many areas, the demand for ever more variables to be extracted from the clinical narrative is currently bottlenecked by the limited supply of technical skills[2]. For example, rule based approaches to novel problems are often effective, but require a certain degree of technical knowledge and experience, which can be too time-consuming and thus expensive to produce in high volume. Proposed solutions include ontological or dictionary mapping techniques, which are appropriate where there are well-constructed resources; however, the standards imposed by these may not be easy to adapt to real world clinical sub-languages[3], or where there is controversy about the appropriate use of clinical language[4]. Machine Learning (ML) approaches are an increasingly popular means of circumventing rule based systems, but require even more technical expertise and are limited by the availability and ease of creating appropriate training data[5,6]. Finally, although progress has been made in the development of publicly available corpora for evaluating different clinical NLP methodologies[7], these offer no guarantee that the performance obtained by models trained on such data will provide a generalizable solution (for example, for work on EHRs in different medical domains, dialects, languages, or work cultures)[8].

These issues form barriers to progress for groups who have access to unstructured clinical data, but do not have sufficient technical capabilities to trial the wealth of information extraction techniques on offer. In recent years this has prompted the development of tools such as Arc[9] to democratize access to generic information extraction capabilities. However, there are currently no free tools available that offer a full end-to-end solution for concept level extraction, including the principle tasks of:

1) Extracting instances of concepts from a database or large collection of documents

2) Creating sufficient training data specific to a concept to enable a machine learning approach

3) The configuration and testing of an (ML) algorithm for the given concept

4) The application of the model to the entire document set of interest, and the subsequent export of results into a familiar format

In order to make concept extraction technologies accessible to groups without informatics support, we have developed the TextHunter tool to address these tasks.

Methods

Data: The South London and Maudsley mental health case register

The South London and Maudsley NHS Trust (SLAM) is the largest mental health organization in Europe, and is a virtual monopoly provider of mental health services to 1.2 million individuals within its geographical catchment area (Lambeth, Southwark, Lewisham and Croydon boroughs in South London). In 2007-08, funding from the British National Institute for Health Research supported the development of the Clinical Record Interactive Search (CRIS) database. CRIS operates as a pseudonymized version of SLAM's EHR system, accessible for researchers via its distinctive, patient-led information governance model[10]. CRIS houses more than 230,000 de-identified patient records, which in turn represent over 20 million free text documents. The CRIS system continues to grow at a rate of approximately 170,000 free text documents per month. Clinical information documented in unstructured text is of particular value in mental health research where there is an increasing emphasis on using dimensional symptom scales to define mental illness rather than discrete diagnostic categories[11–14]. While CRIS also has large amounts of data contained within structured fields, the development of TextHunter was precipitated by the needs of many disparate groups of researchers who require access to the wealth of additional information contained within the clinical narrative.

TextHunter System Description

TextHunter is a program that guides a user through all of the required processes to create and apply a concept extraction model for a selection of documents from start to finish. It performs six important tasks, the end result of which delivers a structured representation of a concept. Its intended use case is typically phenotype cohort identification, although it can be employed for more generic purposes. The program is built from open source libraries, and uses the GATE library as its core NLP engine[15]. The ML element uses the Support Vector Machine (SVM) based 'Batch Learning' plugin supplied with GATE[16]. In consideration of the rigorous information governance requirements of clinical data, TextHunter is designed to operate as a standalone 'offline' program on desktop hardware, although its multithreaded design enables its deployment on more powerful workstations/virtual machines to handle larger datasets. It is capable of connecting to commercial database environments such Microsoft SQL Server to process massive datasets, but for succinctness, only its standalone operation mode is described here.

The underlying principle of TextHunter is 'Find, Annotate, Build, Apply' - respectively addressing the four key problems described above. The integration of these concepts into a single system creates the possibility of providing lay users with access to more advanced ML techniques, such as active learning. Each phase of the TextHunter pipeline is described below:

1. Search Phase

This phase addresses task 1). The first stage of the TextHunter pipeline requires a user to define a list of keywords, regular expressions and/or phrases to describe their concept of interest. The user then directs the program to a directory holding the text files of interest. Upon executing the 'search' phase, each document is scanned for mentions of the user's expressions. When a mention is identified, a short section of text consisting of multiple sentences, including the sentence where the concept mention was found, and up to two sentences either side of the sentence of interest is extracted. This is stored in an embedded file based database, along with a copy of the underlying document. Deconstructing documents in this way facilitates the downstream management of text instances for annotation and classification.

2. Annotation Phase

This phase addresses task 2). The user is directed to TextHunter's annotation interface, which has been specifically designed for the rapid annotation of concept instances. We define an instance as a group of one to five sentences centered on a concept keyword, and its classification as defined below:

i) Positive – the example is a relevant hit and is an appropriate positive example of the user's concept

ii) Negative - the example is a relevant hit and is an appropriate negated example of the user's concept

iii) Unknown - the example is a relevant hit but the user is unable to ascertain the correct classification, or the example is irrelevant

In this phase, the user is required to produce a 'test' corpus for model validation (typically of 100-300 instances), which are randomly selected from all instances in the document set. This is followed by the production of a 'seed' corpus to be used in training models. This also numbers about 100-300 instances, but is enriched by ensuring no identical instances are present. In real world clinical datasets, the required semantic context that enables the classification of a concept instance may cross sentence boundaries. To ensure appropriate features are available for training, the user can specify the required 'context' (up to two sentences before and two sentences after) needed to make the classification, centered on the sentence containing the concept keyword. These boundaries are arbitrarily chosen by the GATE sentence splitter module, although we expect that only in very rare cases will more than five sentences be required to express medical concepts as they are normally found in EHRs.

3. Feature selection/Model Building Phase

This phase addresses task 3). Here, TextHunter builds and evaluates a range of models against the task, using different features and SVM parameters each time. The default feature vector used by TextHunter is a classic bag of words using part-of-speech tags and token stems from the user specified context around a concept. When applying a model to unseen data, TextHunter creates feature vectors from up to six different combinations of sentences around the sentence containing the concept term. The classification resulting from the feature vector producing the highest overall confidence is chosen as the result. In addition, TextHunter has a modular design that allows developments from the clinical NLP community to be integrated into its core pipeline via GATE creole plugins. Currently, TextHunter takes features of the GATE implementation of the ConText algorithm[17], which uses hand crafted rules to determine whether a concept is negated, temporally irrelevant or refers to a subject other than the patient. Stop word removal is also explored during feature selection.

Cross validation of the training data is used to mitigate the dangers of overfitting the model to a small amount of data. The model producing the best F1 score is taken forward for testing against the human labeled 'test' corpus, which is never used in model training. A range of easy to interpret output files are produced, containing estimates of 'real world' performance the user might expect.

4. Application Phase

This phase addresses task 4). This phase allows the user to apply the best performing model to all instances of text in their dataset, as captured in the search phase. As with the model building phase, combinations of sentences are tested around the concept. The classification that results from the combination with the highest confidence is chosen as the final result. Once this stage is complete, the user may export the output into several formats.

5. Active Learning Phase (optional)

Conceptually, active learning is an iterative process whereby an ML algorithm selects instances that it has difficulty classifying and presents them to a human annotator for labeling. These are then fed back into the model, with the intention that the new model arising will be better at classifying similar, difficult examples. TextHunter supports a 'simple margin' inspired method of active learning[18]. A seed model is constructed from randomly selected instances of text, as described above. This model is then applied to a large sample of the entire population of relevant text instances. For each classification the model makes, it also assigns a level of certainty, between -1 and +1. Theoretically, highly positive scores are representative of easy to classify 'positive' instances, whereas highly negative scores are representative of easy to classify 'negative' or 'unknown' instances. Instances with a certainty score close to 0 are thus 'difficult', and presented to the user for labeling in order to retrain the classifier.

Use cases

To evaluate the performance of TextHunter, we defined three real world use cases of concept extraction. Examples of search expressions and typical instances for each use case are detailed in Table 1:

Case Study 1: Cannabis Smoking

Cannabis use has been indicated as a potentially aggravating factor in patients suffering from mental illness[19]. Through the vast amount of electronic documentation generated in the course of patient care, we attempted to identify a patient's cannabis smoking status based upon reports by mental health professionals. The CRIS database contains intra-profession clinical correspondence style documents and clinical notes resulting from patient contact. Each type of document may contain references to cannabis usage by the patient. In this study, our objective was to use TextHunter to build a classifier to identify current or historical cannabis usage. We conducted a review of the most common nouns and slang terms used to describe cannabis in SLAM, to produce a list of expressions which formed the basis for finding instances to classify. A psychiatrist then produced multiple sets of annotations using the standard TextHunter procedure, making use of the active learning functionality. Although it was not possible to double annotate the training data, we adopted a restrictive manual coding strategy in order to allow as little subjectivity as possible (for example, by classifying mentions pertaining to future events, or tangential/circumstantial references into our predefined 'unknown' class).

Case Study 2: Psychosis Symptomatology

Patients suffering from psychosis can exhibit a wide range of symptoms, which in turn inform the nature of their treatment plan. Common tools to quantify symptomatology in psychosis include such instruments as the Positive and Negative Symptom Scale and the Clinical Assessment Interview for Negative Symptoms[12,13]. These depend on an assessment of the patient's presentation in regard to a wide range of possible symptoms. Our previous work to capture some of these from clinical notes with ML approaches has been described[20,21]. In this case study, we used TextHunter to capture two additional symptoms: delusional symptoms and evidence of hallucinations, using the standard TextHunter workflow. The annotated data for 'delusions' were generated by a clinical informatician, with a random sample checked for accuracy and consistency by a psychiatrist. In the case of 'hallucinations', all annotations were generated by a public health physician. In both cases, the restrictive coding strategy as described above was employed.

Table 1: Search expressions and examples of instances. Theoretical patient identifiers masked by ZZZZZ.

Case Study	Examples of subword patterns (case insensitive) for search phase	Fictitious examples of instances (Parentheses indicates typical labeling by human annotator)
Cannabis smoking	cannab hash weed pot	ZZZZZ told me that he continues to smoke cannabis only no other illicit drugs. **(positive)** ZZZZZ has no history of amphetamine or cannabis use. **(negative)**
Psychosis symptomatology	delusio hallucina	She is continuing to experience hallucinations and is becoming increasingly distressed by these. **(positive)** Staff observed him to rambling and delusional, repeating himself and his gait was abnormal and more pronounced. **(positive)**

Case Study 3: Ethnicity

Ethnicity is a key variable in many epidemiological and clinical studies. Although ethnicity can theoretically be captured via the structured elements in SLAM's EHR system, in reality, it is often not recorded in the course of routine clinical practice. However, as with many other variables, ethnicity is often referenced in clinical free text. The purpose of this case study was therefore to classify instances of text describing a patient's ethnicity, into one of 17 ethnic groups. A range of terms was selected in association with each ethnic group, and a 'positive' classification was made if the context for the term was suggestive of the patient belonging to that group. A single researcher produced the annotated dataset for training/testing, using a similarly restrictive coding strategy.

In each case study, a sample of the evaluation instances were double annotated by an individual in a related profession to generate inter-annotator agreement statistics.

Results

In all case studies, we used 10 fold cross validation for the model building phase, which took approximately one hour on a desktop computer with a Core 2 Duo E7500 processor.

In the cannabis smoking study, we used 13 terms to capture cannabis mentions. The CRIS database yielded 663,979 mentions of cannabis. For the psychosis symptomatology study, the search phase found 603,818 mentions of delusions, and 703,996 mentions of hallucinations. Each symptom was represented by a single term in the search phase. Finally, there were 3,444,435 mentions of concepts potentially related to ethnicity, resulting from 277 terms commonly used to define our 17 ethnic identities.

Traditionally, the performances of information extraction algorithms in NLP are described in terms of precision, recall and the F1 statistic. However, the high level of noise commonly associated with EHR based observational research necessitates the capture of high quality data in order to generate clearly defined cohorts. This data quality requirement restricts the use of automated concept extraction techniques to those that can be shown to have a high true positive rate, relative to the inherent predictive value of a mention of a concept. For example, a mention of a cannabis synonym will refer to a patient's current or past use 70% of the time, whereas a mention of a term denoting ethnicity will refer to a patient's actual ethnicity only 20% of the time (Table 1). A further consideration of the real world viability of a given model is the longitudinal nature of the electronic health record. A patient may have numerous contacts with a health service over a number of years, creating multiple instances of time independent concepts. For example, a patient may have multiple references to their cannabis consumption habits, especially if it is identified as a factor in their illness. Similarly, a patient's ethnicity may be described in service referral letters generated during the course of their care. Only one positive instance needs to be captured precisely for a high quality output to be achieved. However, spurious data points are more problematic. Given these factors, it is more practical to develop information extraction tools that favor precision over recall in most use cases. For this reason, in Table 2 we describe the recall statistic at two arbitrarily defined levels of precision (90% and 95%), which are identified by filtering the classified instances in the test set via the classification confidence threshold. We present Receiver-Operator Characteristic (ROC) plots for each case study in Figure 1. For brevity, we only report the highest F1 achieved without any confidence filtering (note, this is not necessarily the same model that achieves the highest recall at the 90%/95% precision threshold).

The best performance was seen in the hallucinations case study, with over 97 % recall obtained at the 95% precision threshold. The worst performance was observed in the ethnicity study, where recall reached only 9% at 90% precision, and declined with further training.

Different problems required different features in order to obtain the best overall result. In Table 3, we present the types of features that were found to be most useful in each case study.

The rate of training data production varied moderately between the studies, the slowest recorded at approximately 100 instances labeled per hour, and the fastest at roughly 230 instances per hour. Since different individuals annotated each study, further comparisons were not possible. Anecdotal reports from the annotators suggested that the process of annotating instances selected via active learning was slower than the randomly selected instances in the seed set.

Table 2: Performance statistics for TextHunter 'positive' instances ('unknown' and 'negative' instances are grouped together). [1]Observed Agreement and Cohen's Kappa. [2]Baseline precision assumes presence of keyword is a 'positive' instance (by definition, recall is 100%), and provides a measure of how predictive a mention is of a concept without any processing applied. P = precision, R = recall, F1 = harmonic mean of precision and recall. [3]Parentheses indicate count of training instances in the model building phase (subsequent active learning iterations increase the number of training instances available). [4]Recall measured at precision levels of 90% and 95%, attained by confidence filtering.

Case Study	Inter annotator agreement [1,3]	Test Instances	Baseline precision[2]	Seed data base performance[3]	Seed data Recall[3,4]		Active learning iteration 1 recall[3,4]		Active learning iteration 2 recall[3,4]		Approximate total annotator time spent creating training data
					90	95	90	95	90	95	
Cannabis smoking	88% 0.76 (211)	233	75%	P = 81% R = 95% F1 = 0.87 (478)	45	38	68	52	72	53	~10 hours
					(478)		(1 329)		(1 835)		
Delusions	95% 0.91 (110)	206	68%	P = 89%/ R = 99%/ F1 = 0.93 (708)	95	87	N/A		N/A		~4 hours
					(708)						
Hallucinations	89% 0.78 (117)	131	70%	P = 93% R = 99% F1 = 0.96 (150)	99	97	99	97	N/A		~7 hours
					(150)		(914)				
Ethnicity	97% 0.94 (201)	650	20%	P = 82% R = 75% F1 = 0.78 (396)	9	9	3	3	N/A		~3 hours
					(396)		(805)				

Table 3: Additional features used in best performing model delivering >90% precision

Case Study	Best Model ID	ConText used?	Stop words removed?	SVM Cost	SVM kernel type
Cannabis Smoking	128	No	No	0.6	polynomial
Delusions	136	No	No	0.6	polynomial
Hallucinations	88	Yes	No	0.5	polynomial
Ethnicity	24	Yes	Yes	0.7	polynomial

Discussion:

In our analysis, we used TextHunter to extract a diverse set of concepts that are typically in demand in clinical research environments. We arbitrarily set two desired precision standards, and adopted strategies to try to maximize the recall given this requirement. Three of the four test cases reached over 70% recall at the lower precision cut-off of 90%. We do not attempt to tackle the question of what constitutes acceptable performance for research applications here. Nevertheless, we have confidence that the range of case studies investigated here establishes a proof of concept in enabling end users to create and deliver information extraction solutions independently of significant NLP expertise.

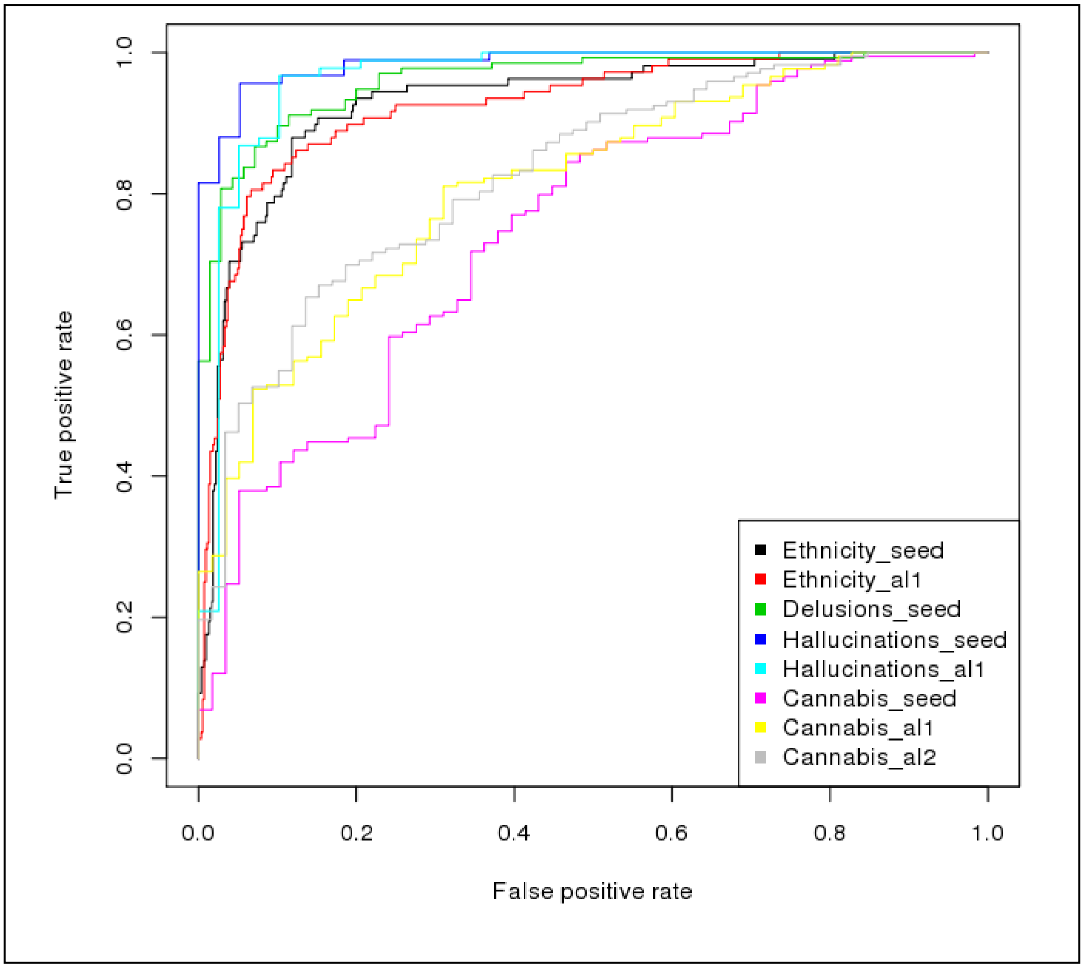


Figure 1: Receiver Operator Characteristic for TextHunter models on 'test' data, generated with SVM confidence thresholds.

Given our limited range of test cases, the SVM parameters and additional features used varied greatly, even between the two conceptually similar problems explored in psychosis symptomatology. This substantiates our approach of testing a range of models to find the best solution for a given problem. However, a predominant factor in the algorithms' ability to reach higher levels of recall is the predictive value that a simple mention of a concept produces (i.e. how likely a human annotator is to label a randomly selected mention of a concept as 'positive'). For instance, the ROC curve produced for the ethnicity study compares favorably with that of the cannabis study, and we achieved a substantial performance benefit over the baseline precision for our list of ethnicity terms. However, because of our self-imposed requirement of a minimum 90% precision, the recall for ethnicity falls very quickly as this threshold is approached. Intuitively, in high noise datasets where 'positive' mentions of a concept are rarer, the concept extraction problem is significantly more challenging. In addition, the low predictive value of ethnicity terms means the 'positive' class will be less represented than the 'negative' or 'unknown' classes in the model. Currently, TextHunter makes no adjustment for unbalanced classes, and future work could investigate mitigation strategies for this, such as using uneven margins [22]. It should also be noted that we were required to use many more terms to capture mentions of ethnicity, which may be indicative of the inherent difficulty of defining concepts that are largely social constructs.

In the case of the cannabis study, we were able to improve the model substantially by providing additional training data through active learning. We did not try to quantify the added benefit of adopting an active learning methodology over randomly selecting new instances. However, others have previously demonstrated that active learning can accelerate the development of machine learning models in clinical NLP[18,23,24]. Active learning did not produce an additional benefit in the hallucinations case study, although the model resulting from the seed data had already produced a very high F1 statistic. Here, our application of confidence filters was not required, as the performance of the model generated from the seed annotations surpassed our precision requirement of 95%. In the case of ethnicity, adopting an active learning approach noticeably depreciated the quality of the model. To investigate, we conducted a subjective review of the instances that active learning retrieved. This revealed that many were incoherent strings of text, seemingly resulting from jumbled emails, faxes and other malformed documents. Since these were not representative of natural language, their inclusion in training the model possibly introduced more noise than benefit. Previous reports have highlighted the difficulties of applying general NLP tools on clinical text[8,25], and we suspect that this scenario is not uncommon in real world EHR systems. One possible mitigation strategy would be to employ document classification methods to filter out malformed documents and/or a more sophisticated active learning methodology, such that new training data are more representative of the instances of interest. Nevertheless, an SVM approach as implemented in TextHunter appears to be valid for simple concepts that tend be succinctly expressed - for example, if it can be defined with a relatively short list of keywords, is not over-complicated by frequent ungrammatical usage (such as in lists or questionnaire text) and has a baseline precision of at least 60%.

It was not practical to double annotate our training data fully, so we are only able to provide inter-annotator agreement (IAA) statistics for a subset of the total test set in each case study. Despite our limited set, our data suggest relatively high levels of agreement, highlighting a high degree of objectivity in the expression of concepts in clinical text. However, clinical constructs in mental illness are often subtle. Initial reports from annotators in each case study suggested that the annotation process itself influenced their own views on the interpretation of notes created by others. Specifically, the exposure to a wide range of writing styles from other clinicians may introduce unforeseeable subjectivity into the annotation process. Regardless, methods that place subject matter experts (rather than NLP specialists) in the role of defining a concept are likely to be less subjective, as any subjectivity introduced by the annotation process will likely be compounded by attempting to convey the subtleties to a non-expert third party. Any clinical subjectivity may then be mitigated by a process of iterative discussion and re-annotation to produce well defined annotation guidelines. A potentially useful future development of TextHunter may be to incorporate a model of clinical data, such as the Clinical Element Model[26]. This would encourage the re-use of standard definitions of concepts, thus promoting greater interoperability with NLP tools.

A notable shortcoming of the TextHunter methodology was the ethnicity case study, which had the highest Kappa statistic but the lowest F1 score from the seed data. This highlights the divide between human and machine interpretation, and the need for more complex reasoning systems to resolve more difficult problems.

Conclusion

The requirement to develop this software was driven by an imbalance between the demand for concept extraction and the supply of skilled individuals capable of delivering solutions to the needs of researchers. We have shown that it is feasible to package an appropriate suite of tools into a simple interface, and that this enables researchers to produce concept extraction models without input from NLP specialists. TextHunter uses a flexible SVM based algorithm as a generic, user friendly information extraction capability. We have validated the methodology with a variety of typical problems, and produced high precision and relatively high recall models. Although it is not suitable for all tasks, we argue that the 'solve small problems quickly' approach to information extraction is appropriate for many types of variable likely to be of interest to researchers, and offers the attractive advantage of rapidly generating models that have been trained on data sourced from the intended target. Finally, the simple annotation interface enables a rapid annotation process, with labeled data stored in a standard, reusable format. The pipeline style operation of GATE and the open source licence of TextHunter should encourage the future development of additional features to improve performance and expedite its use on more complex NLP problems.

TextHunter is available at https://github.com/RichJackson/TextHunter

Funding/Support Acknowledgement

RJ, RD and RS are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. MB is supported by the BRC Nucleus jointly funded by the Guy's and St Thomas' Trustees and the South London and Maudsley Trustees. RP is supported by a Medical Research Council Clinical Research Training Fellowship. RH is funded by a Medical Research Council (MRC) Population Health Scientist Fellowship. RD and RS are joint last authors on this work.

References

1. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association. 2013 Nov 7;21(2):221–30.

2. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association. 2011 Aug 16;18(5):540–3.

3. Demner-Fushman D, Mork JG, Shooshan SE, Aronson AR. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. Journal of Biomedical Informatics. 2010 Aug;43(4):587–94.

4. Chmielewski M, Bagby RM, Markon K, Ring AJ, Ryder AG. Openness to Experience, Intellect, Schizotypal Personality Disorder, and Psychoticism: Resolving the Controversy. J Pers Disord. 2014 Feb 10;

5. Afzal Z, Schuemie MJ, van Blijderveen JC, Sen EF, Sturkenboom MCJM, Kors JA. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC Med Inform Decis Mak. 2013;13:30.

6. Khor R, Yip W-K, Bressel M, Rose W, Duchesne G, Foroudi F. Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements. Journal of the American Medical Informatics Association. 2013 Aug 6;21(1):27–30.

7. Zhai H, Lingren T, Deleger L, Li Q, Kaiser M, Stoutenborough L, et al. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. J Med Internet Res. 2013;15(4):e73.

8. Patterson O, Hurdle J. Document Clustering of Clinical Narratives: a Systematic Study of Clinical Sublanguages. AMIA Annu Symp Proc. 2011.

9. D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. Journal of the American Medical Informatics Association. 2011 Jun 22;18(5):607–13.

10. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry. 2009;9:51.

11. Adam D. Mental health: On the spectrum. Nature. 2013 Apr 24;496(7446):416–8.

12. Kring AM. The Clinical Assessment Interview for Negative Symptoms (CAINS): Final Development and Validation. American Journal of Psychiatry. 2013 Feb 1;170(2):165.

13. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull. 1987;13(2):261–76.

14. Axelrod BN, Goldman RS, Alphs LD. Validation of the 16-item Negative Symptom Assessment. J Psychiatr Res. 1993 Sep;27(3):253–8.

15. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. Prlic A, editor. PLoS Computational Biology. 2013 Feb 7;9(2):e1002854.

16. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011 Apr 1;2(3):1–27.

17. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics. 2009 Oct;42(5):839–51.

18. Koller D, Tong S. Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research. 2001;2:45–66.

19. Moore TH, Zammit S, Lingford-Hughes A, Barnes TR, Jones PB, Burke M, et al. Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review. The Lancet. 2007 Jul;370(9584):319–28.

20. Gorrell G, Jackson R, Roberts A. Finding Negative Symptoms of Schizophrenia in Patient Records. Proc NLP Med Biol Work (NLPMedBio). Hissar, Bulgaria; 2013. p. 9–17.

21. Patel R, Jayatilleke N, Jackson R, Stewart R, McGuire P. Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach. The Lancet. 2014 Feb;383:S16.

22. Li Y, Bontcheva K, Cunningham H. Adapting SVM for data sparseness and imbalance: a case study in information extraction. Natural Language Engineering. 2008 Dec 18;15(02):241.

23. Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, Denny JC, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. Journal of the American Medical Informatics Association. 2013 Jul 13;20(e2):e253–e259.

24. Figueroa RL, Zeng-Treitler Q, Ngo LH, Goryachev S, Wiechmann EP. Active learning for clinical text classification: is it better than random sampling? Journal of the American Medical Informatics Association. 2012 Jun 15;19(5):809–16.

25. Barrett N, Weber-Jahnke JH. Applying natural language processing toolkits to electronic health records - an experience report. Stud Health Technol Inform. 2009;143:441–6.

26. Wu ST, Kaggal VC, Dligach D, Masanz JJ, Chen P, Becker L, et al. A common type system for clinical natural language processing. Journal of Biomedical Semantics. 2013;4(1):1.

7. Conclusions

7.1 Summary of key findings

Three studies were performed using data obtained from the SLaM BRC Case Register in order to investigate clinical outcomes in psychotic disorders relating to high risk clinical services (chapter 3), cannabis use (chapter 4) and negative symptoms (chapter 5).

Chapter 3 demonstrates that presentation with FEP to a high risk service is associated with better clinical outcomes (as indexed by reduced number of inpatient days, reduced frequency of hospital admission and reduced likelihood of compulsory admission) when compared to other clinical services. Around one third of patients presenting to a high risk service are found to already be experiencing their first episode of psychosis.[6,122] Prior to this study, little was known about the future clinical outcomes of this group. The findings in chapter 3 suggest that high risk services may not only support people who are vulnerable to psychosis, but may also play an important role in improving outcomes in people who have already developed a psychotic disorder.

The findings in chapter 4 suggest that a history of cannabis use is associated with a greater frequency of psychiatric hospital admissions, increased likelihood of compulsory hospital admissions and greater number of inpatient days in the five years following presentation to an early intervention service with FEP. Furthermore, these associations were partly mediated by an increase in the number of unique antipsychotics prescribed. Previous studies suggest that cannabis is associated with an increased risk of developing a psychotic disorder.[50,51] However, prior to this study, little was known about the association of cannabis with clinical outcomes in people with an established psychotic disorder. The findings in chapter 4 suggest that cannabis use is associated with worse clinical outcomes in people with FEP which are partly mediated through a failure of antipsychotic treatment.

Chapter 5 demonstrates that negative symptoms are frequently present in people with schizophrenia and that they are associated with increased likelihood of hospital admission, readmission and number of inpatient days. Previous studies investigating negative symptoms in psychosis have involved relatively small research samples using detailed clinical questionnaires which may not be practical to apply in the clinical setting.[123] The findings from chapter 5, which are based on a large sample of patients receiving mental healthcare in a real-world setting, indicate an important role of negative symptoms in predicting poor clinical outcomes.

7.2 Strengths and limitations

The studies reported in this thesis were based on largescale analysis of clinical data from EHRs. A key strength of this approach is the availability of large volumes of data, thereby increasing statistical power to address research questions which would have otherwise been unfeasible to investigate in observational or interventional studies involving direct patient recruitment.[65,66] For example, it would not be feasible to perform a prospective interventional study to investigate the effect of a high risk service on clinical outcomes in people with FEP. The large sample size available in an electronic case register derived from EHRs is coupled with the availability of rich clinical data including free text documents from clinical assessments and correspondence. These documents contain detailed clinical information including the nature of presenting symptoms which would not normally be available in an electronic case register derived from administrative healthcare data. In chapters 4 and 5, the combination of NLP techniques with large volumes of EHR data permitted analysis of data on cannabis use and negative symptoms from large numbers of patients which would otherwise have been difficult to obtain by face-to-face clinical interview or by manually reading clinical documents.

Another strength of EHR data is that they represent naturalistic data collected in the course of providing standard clinical care to patients within a mental healthcare service. This means that the findings drawn from these data are generalizable to the population that normally receives healthcare from mental health services. This is in contrast to data obtained from direct recruitment to observational and interventional studies, when strict inclusion criteria and specialised clinical assessment and intervention can result in findings which are not representative of standard clinical practice.[124] A further benefit of EHR data is that they are less prone to observation bias than data from a study involving direct recruitment.[125] At the time when they were documenting clinical data in EHRs, clinicians were agnostic to the possibility that their data may later be used for research purposes. In particular, in the study reported in chapter 5, clinicians were not specifically seeking to elicit the presence or absence of negative symptoms. This may have resulted in a reduced prevalence of negative symptoms in the SLaM BRC Case Register data compared to previous studies involving specialised research instruments specifically designed to obtain data on negative symptoms. However, the benefit of analysing EHR data to investigate negative symptoms is that clinicians were unbiased with respect to the future use of their records for this purpose and so their assessments in EHRs may be more representative of the prevalence of negative symptoms elicited in standard clinical practice.

The primary limitation in drawing conclusions from the EHR data is that they are observational, and so any association of a predictor with subsequent clinical outcomes does not necessarily demonstrate an aetiological association. Another difficulty is that by virtue of being naturalistic, clinical data were not comprehensively collected for all the patients included in the studies. This resulted in some missing covariate data. Sensitivity analyses including missing data as a separate category in multivariable regression analyses did not result in meaningful changes to the findings. However, there were other variables which were not possible to analyse in detail due to a lack of consistent clinical documentation. These included details on medication dose, duration, concordance, side effects and reason for discontinuation (if applicable) which would have helped to better explain the finding that cannabis is associated with an increase in the number of unique antipsychotics prescribed to people with FEP (chapter 4). A furether limitation is lack of documentation on change in nature and degree of symptomatology over time which limited the analysis of negative symptoms in schizophrenia (chapter 5) to the presence and absence of negative symptoms without examining degree or severity of symptoms.

Related to this issue, by virtue of the fact that clinical data are only documented in EHRs when patients receive care from clinical services. This means that on the balance of probability, NLP techniques are likely to obtain more data from patients who have a greater degree and frequency of contact with mental health service. It is possible that the absence of documentation on a patient could have been biased by the fact that they were well, and so did not require mental healthcare services. Conversely, the absence of documentation could reflect disengagement from mental health services by people who are unwell or experiencing a relapse of psychotic symptoms. Therefore, NLP techniques may have underestimated the presence of cannabis use (chapter 4) or negative symptoms (chapter 5) in patients who had less contact with mental health services by virtue of having been thought to be well enough not to require these services or to be unwell because of disengagement from mental health services.

A further limitation in EHR data is the degree to which constructs documented in clinical records represent the clinical presentation to which they refer within individual patients and the degree to which they are comparable between different patients. It is not possible to know whether different clinicians (with different levels of expertise and experience) who assessed the same patients would have reported the same findings. Nonetheless, it is likely that if a feature of a patient's history or mental state is particularly relevant to their clinical presentation (e.g. if it influences subsequent treatment decisions and management), it is likely to have been documented in the course of their clinical care. Likewise, patients are often observed and assessed by a range of different mental healthcare professionals including doctors, nurses, psychologists and other allied healthcare professionals. This would increase the likelihood that relevant presenting features are documented at some point in a patient's EHR.

7.3 Application of CRIS and NLP methods in other healthcare centres

The widespread use of EHR systems in mental healthcare services raises the possibility of applying the data extraction methods described in this thesis in other healthcare centres. Following the implementation of CRIS in SLaM, a collaboration has developed to implement the CRIS software in five providers of mental healthcare in Southeast England: SLaM, Camden and Islington, West London Mental Health, Cambridgeshire and Peterborough and Oxford Health NHS Trusts.[126] This collaboration demonstrates the potential for the integration of a data extraction and assembly framework based on Microsoft SQL Server (i.e. the CRIS tool) to be implemented in different healthcare settings using different EHR systems.

However, the generalisability of CRIS to other EHR systems depends on the data structure and framework used to store clinical data. While the composition of structured text fields may vary between different EHR systems, those storing core clinical data such as age, gender, ethnicity and other sociodemographic data are likely to be present among all systems. Beyond structured fields, free text clinical data from assessments and correspondence stored in unstructured fields are also likely to be present as free text records are the most widely used to store clinical data.[98] It is likely that the NLP approaches employed in this thesis could be applied to free text clinical records from other healthcare centres. However, the extent to which NLP algorithms derived in one centre could be directly translated to another EHR system remains uncertain.

It is possible that local variations in the documentation of clinical information may affect the ability of NLP software derived in one healthcare centre to accurately identify data in another centre, even if it is written in the same human language. This may be due to local variations in vocabulary used by clinicians to describe clinical phenotypes, differences in spelling and grammar or variation in the vocabulary used by patients to describe their clinical presentation.[127] However, within the UK, these differences are unlikely to be large.

One method of overcoming this limitation may be to derive NLP algorithms which are unique to individual healthcare providers. The same key words used to derive the NLP algorithms to identify cannabis use and negative symptoms in this thesis could be used to generate a reference and training dataset using free text clinical data from an EHR system in another healthcare centre. This training data could then be used to generate an NLP algorithm which is specific to the EHR system in which it is to be applied. Further work is needed to evaluate the potential for this approach or whether NLP algorithms derived in EHR system could be directly applied in another.

7.4 Future research

The studies reported in this thesis demonstrate the potential for EHR data (supplemented by NLP) to contribute to mental health research. In these studies, prospectively recorded clinical data were analysed in retrospect to examine the impact of high risk clinical services, cannabis and negative symptoms on clinical outcomes. The findings from these studies prompt further research with two aims: firstly, to address the limitations described in section 6.2 in order to improve the method and secondly, to consider how the methods developed as a consequence of this work could be used for secondary purposes.

The extent to which data obtained from EHRs using NLP represent an accurate portrayal of the clinical presentation of an individual patient (i.e. construct validity) remains unclear. In order to address this limitation, future studies may benefit from combining data obtained using NLP with data obtained using structured diagnostic interviews which have already been validated in several research samples.[43,128] For example, one could compare the results of an NLP assessment of negative symptoms in EHRs with data obtained from the direct evaluation of negative symptoms in the same patient sample, using psychopathological rating scales. Similarly, clinical outcomes in FEP patients who had initially presented to different types of service could be assessed through follow-up interviews and compared with the outcomes derived from the application of CRIS to the samples' EHRs.

A further consideration is the heterogeneity between volume and quality of clinical documentation in different patients, which introduces heterogeneity into the data that NLP methods are used to examine. Clinical entries can be made by clinicians from a variety of different health professions (e.g. doctors, nurses, occupational therapists, psychologists), and a variety of different people within each profession. Most of the information recorded is in unstructured clinical documents, and the frequency of entries, and the volume of information is uncontrolled. In the studies reported in this thesis, and in the literature more generally, no attempt has been made to standardise the analyses to take this variability into account. One approach to address this issue would be to investigate methods of standardising EHR data prior to the application of NLP analyses. This could include adjusting for total volume of clinical documentation per patient, or the development of additional NLP applications to automatically identify particular types of clinical document (such as discharge summaries and outpatient clinical letters) that are likely to yield the most relevant and high-quality data. This approach may help to improve the precision of NLP data extraction. An alternative approach may be to develop methods to improve standardisation of data entry by means of structured text fields. The use of NLP on clinical assessment data recorded in unstructured free text could be used as a screening tool to highlight clinical features which are relevant to a patient's clinical presentation and prompt clinicians to complete a structured assessment tool. For example, NLP applications to ascertain negative symptoms (described in chapter 5) could be applied in realtime to an individual patient's clinical record and prompt clinicians to consider completing a standardised assessment tool such as the Brief Negative Symptoms Scale (BNSS)[43] or the Scale for the Assessment of Negative Symptoms (SANS).[44] This would allow for NLP applications to support more focussed clinical assessments which are tailored towards a patient's individual clinical presentation.

In addition to epidemiological research using historical EHR data, the data extraction methods developed in this thesis have the potential to be applied to live EHR data for the purposes of supporting clinical decision-making. Even at the individual patient level, the volume of clinical documentation for a patient with a long history of contact with mental health services (possibly several hundreds of pages in length) can make it unfeasible for clinicians to review the entire unstructured free text record prior to assessing or reviewing a patient. NLP offers the opportunity to automatically extract clinically useful information which could be fed back to clinicians to help support clinical assessment and treatment planning in real-time. In order to facilitate this, future research would need to focus on developing additional NLP applications to extract the presence of other symptoms relevant to patients with psychotic disorders (e.g. positive, affective and cognitive symptoms) as well as better delineating the evolution of clinical parameters identified using NLP (e.g. cannabis use or negative symptoms) over time.

A further potential is for automated data extraction tools to support automated clinical prediction tools. One of the challenges faced by clinicians in treating individuals with psychotic disorders is that it is currently not possible to predict prognosis or response to treatment. Between 20-40% of people who develop a psychotic disorder are admitted to hospital in the first year following presentation to mental health services with a mean duration of admission of 2 months.[129] However, at present, it is not possible to predict how long an individual patient is likely to spend in hospital. Similarly, about a third of patients with psychosis do not respond well to conventional antipsychotic medication. Again, this is not predictable on the basis of their presenting clinical features, and can only be determined through a lengthy processing of trial and error, with the evaluation of a series of different medications over several weeks. Future studies could explore the possibility of using structured and unstructured EHR data to develop tools to predict the likely duration of hospital admission or the response to treatment. Such an approach may allow patients and their carers to better plan their recovery and future as well as reducing anxiety surrounding the uncertainty of their illness and may facilitate clinical decision making. The identification of patients who are unlikely to respond to conventional treatment would allow them to be offered alternative treatments (such as clozapine) at a much earlier stage. There are inherent challenges to this approach related to whether clinical prediction tools derived from one sample could be generalised to another sample either in the same healthcare provider or in another provider. Previous studies attempting to apply machine learning techniques to neuroimaging data for the purposes of individual clinical prediction have highlighted limitations in the ability of such tools to make predictions outside of the research sample from which they were derived.[117] However, it is thought that by combining multimodal data (e.g.

genetics, neuroimaging, serum biomarkers), it may be possible to overcome this limitation. A number of ongoing studies including EU-GEI,[130] OPTIMISE,[131] PSYSCAN[132] and STRATA[133] are investigating the possibility of combining multimodal data to make predictions on risk of developing psychosis and likelihood of response to antipsychotic treatment.

7.5 Summary

The application of data extraction methods (including NLP) to obtain EHR data provides an opportunity to conduct clinical research in samples of patients that are much larger than can be directly recruited to conventional research projects. Using these methods, I have demonstrated that people with FEP who present to high risk services have better clinical outcomes than those who present to other services, that people with FEP who use cannabis have worse clinical outcomes than those who do not, and that negative symptoms in people with schizophrenia are common and are associated with particularly poor outcomes. The methods developed in these studies have the potential to instruct automated clinical prediction to support clinical decision making at an individual patient level.

References

1 Perälä J, Suvisaari J, SI S, *et al.* LIfetime prevalence of psychotic and bipolar i disorders in a general population. *Arch Gen Psychiatry* 2007;**64**:19–28.http://dx.doi.org/10.1001/archpsyc.64.1.19

2 McCrone P, Dhanasiri S, Patel A, *et al. Paying the Price: The cost of mental health care in England to 2026*. The King's Fund 2008.

3 Van Os J, Marcelis M, Sham P, *et al.* Psychopathological syndromes and familial morbid risk of psychosis. *Br J Psychiatry* 1997;**170**:241–6. doi:10.1192/bjp.170.3.241

4 Insel TR, Cuthbert BN, Garvey MA, *et al.* Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry* 2010;**167**:748–51.

5 Fusar-Poli P, Borgwardt S, Bechdolf A, *et al.* The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry* 2013;**70**:107–20.

6 Fusar-Poli P, Byrne M, Badger S, *et al.* Outreach and support in South London (OASIS), 2001–2011: ten years of early diagnosis and treatment for young individuals at high clinical risk for psychosis. *Eur Psychiatry* 2013;**28**:315–26.

7 Jääskeläinen E, Juola P, Hirvonen N, *et al.* A systematic review and meta-analysis of recovery in schizophrenia. *Schizophr Bull* 2012;:sbs130.

8 Morgan C, Lappin J, Heslin M, *et al.* Reappraising the long-term course and outcome of psychotic disorders: the AESOP-10 study. *Psychol Med* 2014;**44**:2713–26.

9 Lindenmayer J-P. Treatment refractory schizophrenia. *Psychiatr Q* 2000;**71**:373–84.

10 Clemmensen L, Vernal DL, Steinhausen H-C. A systematic review of the long-term outcome of early onset schizophrenia. *BMC Psychiatry* 2012;**12**:150.

11 Demjaha a, Morgan K, Morgan C, *et al.* Combining dimensional and categorical representation of psychosis: the way forward for DSM-V and ICD-11? *Psychol Med* 2009;**39**:1943–55. doi:10.1017/S0033291709990651

12 Guerra A, Fearon P, Sham P, *et al.* The relationship between predisposing factors, premorbid function and symptom dimensions in psychosis: an integrated approach. *Eur Psychiatry* 2002;**17**:311–20. doi:http://dx.doi.org/10.1016/S0924-9338(02)00685-5

13 Cuesta MJ, Peralta V. Integrating psychopathological dimensions in functional psychoses: a hierarchical approach. *Schizophr Res* 2001;**52**:215–29. doi:http://dx.doi.org/10.1016/S0920-9964(00)00190-0

14 Allardyce J, McCreadie RG, Morrison G, *et al.* Do symptom dimensions or categorical diagnoses best discriminate between known risk factors for psychosis? *Soc Psychiatry Psychiatr Epidemiol* 2007;**42**:429–37. doi:10.1007/s00127-007-0179-y

15 White C, Stirling J, Hopkins R, *et al.* Predictors of 10-year outcome of first-episode psychosis. *Psychol Med* 2009;**39**:1447–56. doi:10.1017/S003329170800514X

16 Jager M, Riedel M, Schmauss M, *et al.* Prediction of symptom remission in schizophrenia during inpatient treatment. World J. Biol. Psychiatry. 2009;**10**:426–34.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=17853260

17 Hunter R, Barry S. Negative symptoms and psychosocial functioning in schizophrenia: neglected but important targets for treatment. Eur. Psychiatry. 2012;**27**:432–6.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=21602034

18 Malla AK, Norman RMG, Takhar J, *et al.* Can patients at risk for persistent negative symptoms be identified during their first episode of psychosis?. J. Nerv. Ment. Dis. 2004;**192**:455–63.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=15232315

19 Ucok A, Serbest S, Kandemir PE. Remission after first-episode schizophrenia: results of a long-term follow-up. Psychiatry Res. 2011;**189**:33–7.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=21196051

20 Tandon R, DeQuardo JR, Taylor SF, *et al.* Phasic and enduring negative symptoms in schizophrenia: biological markers and relationship to outcome. Schizophr. Res. 2000;**45**:191–201.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=11042437

21 Moller HJ, Bottlender R, Wegner U, *et al.* Long-term course of schizophrenic, affective and schizoaffective psychosis: focus on negative symptoms and their impact on global indicators of outcome. Acta Psychiatr. Scand. Suppl. 2000;:54–7.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=11261641

22 Dominguez M-G, Saka MC, can Saka M, *et al.* Early expression of negative/disorganized symptoms predicting psychotic experiences and subsequent clinical psychosis: a 10-year study. *Am J Psychiatry* 2010;**167**:1075–82. doi:10.1176/appi.ajp.2010.09060883

23 Dominguez MDG, Wichers M, Lieb R, *et al.* Evidence that onset of clinical psychosis is an outcome of progressively more persistent subclinical psychotic experiences: an 8-year cohort study. *Schizophr Bull* 2011;**37**:84–93. doi:10.1093/schbul/sbp022

24 McGurk SR, Moriarty PJ, Harvey PD, *et al.* Relationship of cognitive functioning, adaptive life skills, and negative symptom severity in poor-outcome geriatric schizophrenia patients. J. Neuropsychiatry Clin. Neurosci. 2000;**12**:257–64.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=11001606

25 Kimhy D, Yale S, Goetz RR, *et al.* The Factorial Structure of the Schedule for the Deficit Syndrome in Schizophrenia. *Schizophr Bull* 2006;**32** :274–8. doi:10.1093/schbul/sbi064

26 Strauss GP, Horan WP, Kirkpatrick B, *et al.* Deconstructing negative symptoms of schizophrenia: avolition-apathy and diminished expression clusters predict clinical presentation and functional outcome. *J Psychiatr Res* 2013;**47**:783–90. doi:10.1016/j.jpsychires.2013.01.015

27 Salamone JD, Koychev I, Correa M, *et al.* Neurobiological basis of motivational deficits in psychopathology. *Eur Neuropsychopharmacol* 2014.

28 Sigmundsson T, Suckling J, Maier M, *et al.* Structural abnormalities in frontal, temporal, and limbic regions and interconnecting white matter tracts in schizophrenic patients with prominent negative symptoms. *Am J Psychiatry* 2001;**158**:234–43.http://www.ncbi.nlm.nih.gov/pubmed/11156806 (accessed 11 Jul2013).

29 Nakamura K, Kawasaki Y, Takahashi T, *et al.* Reduced white matter fractional anisotropy and clinical symptoms in schizophrenia: a voxel-based diffusion tensor imaging study. *Psychiatry Res* 2012;**202**:233–8. doi:10.1016/j.pscychresns.2011.09.006

30 Goghari VM, Sponheim SR, MacDonald AW. The functional neuroanatomy of symptom dimensions in schizophrenia: a qualitative and quantitative review of a persistent question. *Neurosci Biobehav Rev* 2010;**34**:468–86. doi:10.1016/j.neubiorev.2009.09.004

31 Egerton A, Brugger S, Raffin M, *et al.* Anterior Cingulate Glutamate Levels Related to Clinical

Status Following Treatment in First-Episode Schizophrenia. *Neuropsychopharmacology* 2012;**37**:2515–21.http://dx.doi.org/10.1038/npp.2012.113

32 Fusar-Poli P, Papanastasiou E, Stahl D, *et al.* Treatments of Negative Symptoms in Schizophrenia: Meta-Analysis of 168 Randomized Placebo-Controlled Trials. *Schizophr Bull* 2014;:sbu170.

33 Iancu I, Tschernihovsky E, Bodner E, *et al.* Escitalopram in the treatment of negative symptoms in patients with chronic schizophrenia: a randomized double-blind placebo-controlled trial. Psychiatry Res. 2010;**179**:19–23.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=20472299

34 Singh SP, Singh V, Kar N, *et al.* Efficacy of antidepressants in treating the negative symptoms of chronic schizophrenia: meta-analysis. Br. J. Psychiatry. 2010;**197**:174–9.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=20807960

35 Olie J-P, Spina E, Murray S, *et al.* Ziprasidone and amisulpride effectively treat negative symptoms of schizophrenia: results of a 12-week, double-blind study. Int. Clin. Psychopharmacol. 2006;**21**:143–51.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=16528136

36 Stauffer VL, Song G, Kinon BJ, *et al.* Responses to antipsychotic therapy among patients with schizophrenia or schizoaffective disorder and either predominant or prominent negative symptoms. Schizophr. Res. 2012;**134**:195–201.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=22019076

37 Klingberg S, Wolwer W, Engel C, *et al.* Negative symptoms of schizophrenia as primary target of cognitive behavioral therapy: results of the randomized clinical TONES study. Schizophr. Bull. 2011;**37 Suppl 2**:S98–110.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=21860053

38 Roffman JL, Lamberti JS, Achtyes E, *et al.* Randomized multicenter investigation of folate plus vitamin B12 supplementation in schizophrenia. JAMA Psychiatry. 2013;**70**:481–9.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=23467813

39 Arbabi M, Bagheri M, Rezaei F, *et al.* A placebo-controlled study of the modafinil added to risperidone in chronic schizophrenia. Psychopharmacology (Berl). 2012;**220**:591–8.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=21947320

40 Muller N, Krause D, Dehning S, *et al.* Celecoxib treatment in an early stage of schizophrenia: results of a randomized, double-blind, placebo-controlled trial of celecoxib augmentation of amisulpride treatment. Schizophr. Res. 2010;**121**:118–24.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=20570110

41 Barr MS, Farzan F, Tran LC, *et al.* A randomized controlled trial of sequentially bilateral prefrontal cortex repetitive transcranial magnetic stimulation in the treatment of negative symptoms in schizophrenia. Brain Stimul. 2012;**5**:337–46.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=21782542

42 Arango C, Garibaldi G, Marder SR. Pharmacological approaches to treating negative symptoms: A review of clinical trials. *Schizophr Res* 2013;**150**:346–52. doi:http://dx.doi.org/10.1016/j.schres.2013.07.026

43 Kirkpatrick B, Strauss GP, Nguyen L, *et al.* The brief negative symptom scale: psychometric properties. *Schizophr Bull* 2011;**37**:300–5.

44 NC A. Negative symptoms in schizophrenia: Definition and reliability. *Arch Gen Psychiatry* 1982;**39**:784–8.http://dx.doi.org/10.1001/archpsyc.1982.04290070020005

45 United Nations Office on Drugs and Crime. World Drug Report. 2013. http://www.unodc.org/unodc/secured/wdr/wdr2013/World_Drug_Report_2013.pdf

46 McManus S, Meltzer H, Brugha TS, *et al.* Adult psychiatric morbidity in England, 2007: results of a household survey. 2009.

47 Horwood LJ, Fergusson DM, Hayatbakhsh MR, *et al.* Cannabis use and educational achievement: Findings from three Australasian cohort studies. *Drug Alcohol Depend* 2010;**110**:247–53.

48 Fergusson DM, Horwood LJ, Beautrais AL. Cannabis and educational achievement. *Addiction* 2003;**98**:1681–92.

49 Fergusson DM, Boden JM. Cannabis use and later life outcomes. *Addiction* 2008;**103**:969–76.

50 Moore THM, Zammit S, Lingford-Hughes A, *et al.* Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review. *Lancet* 2007;**370**:319–28.

51 Kuepper R, van Os J, Lieb R, *et al.* Continued cannabis use and risk of incidence and persistence of psychotic symptoms: 10 year follow-up cohort study. *BMJ* 2011;**342**.

52 Arseneault L, Cannon M, Poulton R, *et al.* Cannabis use in adolescence and risk for adult psychosis: longitudinal prospective study. *BMJ* 2002;**325**:1212–3.

53 Di Forti M, Marconi A, Carra E, *et al.* Proportion of patients in south London with first-episode psychosis attributable to use of high potency cannabis: a case-control study. *The Lancet Psychiatry* 2015;**2**:233–8.

54 Bhattacharyya S, Atakan Z, Martin-Santos R, *et al.* Preliminary report of biological basis of sensitivity to the effects of cannabis on psychosis: AKT1 and DAT1 genotype modulates the effects of δ-9-tetrahydrocannabinol on midbrain and striatal function. *Mol Psychiatry* 2012;**17**:1152.

55 Caspi A, Moffitt TE, Cannon M, *et al.* Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene X environment interaction. *Biol Psychiatry* 2005;**57**:1117–27.

56 Foti DJ, Kotov R, Guey LT, *et al.* Cannabis use and the course of schizophrenia: 10-year follow-up after first hospitalization. *Am J Psychiatry* 2010;**167**:987–93.

57 Green AI, Tohen MF, Hamer RM, *et al.* First episode schizophrenia-related psychosis and substance use disorders: acute response to olanzapine and haloperidol. *Schizophr Res* 2004;**66**:125–35.

58 Zammit S, Moore THM, Lingford-Hughes A, *et al.* Effects of cannabis use on outcomes of psychotic disorders: systematic review. *Br J Psychiatry* 2008;**193**:357–63.

59 Hides L, Dawe S, Kavanagh DJ, *et al.* Psychotic symptom and cannabis relapse in recent-onset psychosis. *Br J Psychiatry* 2006;**189**:137–43.

60 Verdoux H, Liraud F, Gonzales B, *et al.* Predictors and outcome characteristics associated with suicidal behaviour in early psychosis: a two-year follow-up of first-admitted subjects. *Acta*

Psychiatr Scand 2001;**103**:347–54.

61 Barrowclough C, Gregg L, Lobban F, *et al.* The Impact of Cannabis Use on Clinical Outcomes in Recent Onset Psychosis. *Schizophr Bull* 2014;:sbu095.

62 Faber G, Smid HG, Van Gool AR, *et al.* Continued cannabis use and outcome in first-episode psychosis: data from a randomized, open-label, controlled trial. *J Clin Psychiatry* 2012;**73**:632–8.

63 van Dijk D, Koeter MWJ, Hijman R, *et al.* Effect of cannabis use on the course of schizophrenia in male patients: a prospective cohort study. *Schizophr Res* 2012;**137**:50–7.

64 Stone JM, Fisher HL, Major B, *et al.* Cannabis use and first-episode psychosis: relationship with manic and psychotic symptoms, and with age at presentation. *Psychol Med* 2014;**44**:499–506.

65 Perera G, Soremekun M, Breen G, *et al.* The psychiatric case register: noble past, challenging present, but exciting future. *Br J Psychiatry* 2009;**195**:191–3.

66 Stewart R. The big case register. *Acta Psychiatr Scand* 2014;**130**:83–6.

67 Cullen R, Caskey F, Fogarty D, *et al.* UK Renal Registry. *Nephron Clin Pract* 2013;**125**:I – XII.http://www.karger.com/DOI/10.1159/000360019

68 Department of Health. National Cancer Registration Service Database. Natl. Cancer Regist. Serv. Engl. 2010.http://www.ncr.nhs.uk/

69 Williams T, Van Staa T, Puri S, *et al.* Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv Drug Saf* 2012;**3**:89–99.

70 Mors O, Perto GP, Mortensen PB. The Danish psychiatric central research register. *Scand J Public Health* 2011;**39**:54–7.

71 Okkels N, Vernal DL, Jensen SOW, *et al.* Changes in the diagnosed incidence of early onset schizophrenia over four decades. *Acta Psychiatr Scand* 2013;**127**:62–8.

72 Reutfors J, Bahmanyar S, Jönsson EG, *et al.* Diagnostic profile and suicide risk in schizophrenia spectrum disorder. *Schizophr Res* 2010;**123**:251–6.

73 Osborn DPJ, Hardoon S, Omar RZ, *et al.* Cardiovascular risk prediction models for people with severe mental illness: results from the prediction and management of cardiovascular risk in People with Severe Mental Illnesses (PRIMROSE) Research Program. *JAMA Psychiatry* 2015;**72**:143–51.

74 Munk-Jørgensen P, Okkels N, Golberg D, *et al.* Fifty years' development and future perspectives of psychiatric register research. *Acta Psychiatr Scand* 2014;**130**:87–98.

75 OVID Gateway. Wolters Kluwer Heal. 2015.http://gateway.ovid.com

76 Taylor CL, Stewart R, Ogden J, *et al.* The characteristics and health needs of pregnant women with schizophrenia compared with bipolar disorder and affective psychoses. *BMC Psychiatry* 2015;**15**:88.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=25886140

77 Nielsen J, Nielsen RE, Correll CU. Predictors of clozapine response in patients with treatment-refractory schizophrenia: results from a Danish Register Study. *J Clin Psychopharmacol* 2012;**32**:678–83.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=22926603

78 Liao Y-T, Yang S-Y, Liu H-C, *et al.* Cardiac complications associated with short-term mortality in schizophrenia patients hospitalized for pneumonia: a nationwide case-control study. *PLoS*

One 2013;**8**:e70142.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=23922940

79 Kuo S-C, Chen Y-T, Li S-Y, *et al.* Incidence and outcome of newly-diagnosed tuberculosis in schizophrenics: a 12-year, nationwide, retrospective longitudinal study. *BMC Infect Dis* 2013;**13**:351.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=23895638

80 Kang J-H, Xirasagar S, Lin H-C. Lower mortality among stroke patients with schizophrenia: a nationwide population-based study. *Psychosom Med* 2011;**73**:106–11.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=20978225

81 Wu J, He X, Liu L, *et al.* Health care resource use and direct medical costs for patients with schizophrenia in Tianjin, People's Republic of China. *Neuropsychiatr Dis Treat* 2015;**11**:983–90.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=25897235

82 Velligan DI, Carroll C, Lage MJ, *et al.* Outcomes of medicaid beneficiaries with schizophrenia receiving clozapine only or antipsychotic combinations. *Psychiatr Serv* 2015;**66**:127–33.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=25321616

83 Hendrie HC, Tu W, Tabbey R, *et al.* Health outcomes and cost of care among older adults with schizophrenia: a 10-year study using medical records across the continuum of care. *Am J Geriatr Psychiatry* 2014;**22**:427–36.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=23933423

84 Schreiner A, Svensson A, Wapenaar R, *et al.* Long-acting injectable risperidone and oral antipsychotics in patients with schizophrenia: results from a prospective, 1-year, non-interventional study (InORS). *World J Biol Psychiatry* 2014;**15**:534–45.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=24779526

85 Williams R, Chandrasena R, Beauclair L, *et al.* Risperidone long-acting injection in the treatment of schizophrenia: 24-month results from the electronic Schizophrenia Treatment Adherence Registry in Canada. *Neuropsychiatr Dis Treat* 2014;**10**:417–25.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=24600227

86 Lambert T, Olivares JM, Peuskens J, *et al.* Effectiveness of injectable risperidone long-acting therapy for schizophrenia: data from the US, Spain, Australia, and Belgium. *Ann Gen Psychiatry* 2011;**10**:10.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=21463526

87 Chan SKW, So HC, Hui CLM, *et al.* 10-year outcome study of an early intervention program for psychosis compared with standard care service. *Psychol Med* 2015;**45**:1181–93.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=25233868

88 Frey S. The economic burden of schizophrenia in Germany: a population-based retrospective cohort study using genetic matching. *Eur Psychiatry* 2014;**29**:479–89.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=24853296

89 Nielsen J, Jensen SOW, Friis RB, *et al.* Comparative effectiveness of risperidone long-acting injectable vs first-generation antipsychotic long-acting injectables in schizophrenia: results from a nationwide, retrospective inception cohort study. *Schizophr Bull* 2015;**41**:627–36.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=25180312

90 Schreiner A, Hargarter L, Hitschfield K, *et al.* Clinical effectiveness and resource utilization of paliperidone ER for schizophrenia: Pharmacoepidemiologic International Longitudinal Antipsychotic Registry (PILAR). *Curr Med Res Opin* 2014;**30**:1279–89.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=24597755

91 Mustafa MZ, Schofield J, Mills PR, *et al.* The efficacy and safety of treating hepatitis C in patients with a diagnosis of schizophrenia. *J Viral Hepat* 2014;**21**:e48–51.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=24533990

92 Ebdrup NH, Assens M, Hougaard CO, *et al.* Assisted reproductive technology (ART) treatment in women with schizophrenia or related psychotic disorder: a national cohort study. *Eur J Obstet Gynecol Reprod Biol* 2014;**177**:115–20.http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=medl&NEWS=N&AN=24721442

93 Stewart R, Soremekun M, Perera G, *et al.* The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* 2009;**9**:51.http://www.biomedcentral.com/1471-244X/9/51

94 Wu C-Y, Chang C-K, Hayes RD, *et al.* Clinical risk assessment rating and all-cause mortality in secondary mental healthcare: the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) Case Register. *Psychol Med* 2012;**42**:1581–90.

95 Chang C-K, Hayes R, Broadbent M, *et al.* All-cause mortality among people with serious mental illness (SMI), substance use disorders, and depressive disorders in southeast London: a cohort study. *BMC Psychiatry* 2010;**10**:77.

96 Hayes RD, Chang C-K, Fernandes A, *et al.* Associations between substance use disorder sub-groups, life expectancy and all-cause mortality in a large British specialist mental healthcare service. *Drug Alcohol Depend* 2011;**118**:56–61. doi:http://dx.doi.org/10.1016/j.drugalcdep.2011.02.021

97 Chang C-K, Hayes RD, Perera G, *et al.* Life Expectancy at Birth for People with Serious Mental Illness and Other Major Disorders from a Secondary Mental Health Care Case Register in London. *PLoS One* 2011;**6**:e19590.http://dx.doi.org/10.1371/journal.pone.0019590

98 Meystre SM, Savova GK, Kipper-Schuler KC, *et al.* Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inf* 2008;**35**:128–44.

99 Cunningham H, Tablan V, Roberts A, *et al.* Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Comput Biol* 2013;**9**:e1002854.http://dx.doi.org/10.1371%2Fjournal.pcbi.1002854

100 Li Y, Bontcheva K, Cunningham H. Adapting SVM for data sparseness and imbalance: a case study in information extraction. *Nat Lang Eng* 2009;**15**:241–71. doi:10.1017/S1351324908004968

101 Murff HJ, FitzHenry F, Matheny ME, *et al.* Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA* 2011;**306**:848–55.

102 Sohn S, Kocher J-PA, Chute CG, *et al.* Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Informatics Assoc* 2011;**18**:i144–9.

103 Perlis RH, Iosifescu D V, Castro VM, *et al.* Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med* 2012;**42**:41–50.

104 Wu C-Y, Chang C-K, Robson D, *et al.* Evaluation of Smoking Status Identification Using Electronic Health Records and Open-Text Information in a Large Mental Health Case Register. *PLoS One* 2013;**8**:e74262.

105 Su Y-P, Chang C-K, Hayes RD, *et al.* Mini-Mental State Examination as a Predictor of Mortality among Older People Referred to Secondary Mental Healthcare. *PLoS One* 2014.

106 Patel R, Lloyd T, Jackson R, *et al.* Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. *BMJ Open* 2015;**5** . doi:10.1136/bmjopen-2014-007504

107 *Understanding the new NHS*. NHS England 2014. http://www.england.nhs.uk/wp-content/uploads/2014/06/simple-nhs-guide.pdf

108 South London and Maudsley NHS Foundation Trust. Strategic Plan 2014-19 summary. 2014. http://www.slam.nhs.uk/media/313512/Strategic Plan 2014-19.pdf

109 South London and Maudsley NHS Foundation Trust. Clinical Academic Groups. 2015.http://www.slam.nhs.uk/about-us/clinical-academic-groups

110 Marshall M, Rathbone J. Early intervention for psychosis. *Cochrane Database Syst Rev* 2011;:CD004718. doi:10.1002/14651858.CD004718.pub3

111 Burns T, Creed F, Fahy T, *et al.* Intensive versus standard case management for severe psychotic illness: a randomised trial. *Lancet* 1999;**353**:2185–9.

112 Nielsen P, Parui U. *Microsoft SQL Server 2008 Bible*. John Wiley & Sons 2011.

113 Fernandes AC, Cloete D, Broadbent MTM, *et al.* Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Med Inform Decis Mak* 2013;**13**:71.

114 Mental Health Act. Great Britain: : London: The Stationery Office 2007. http://www.legislation.gov.uk/ukpga/2007/12/contents

115 *International statistical classification of diseases and related health problems*. 10th revis. World Health Organization 2004.

116 Joint_Formulary_Committee. *British National Formulary*. Pharmaceutical Press 2013.

117 Kempton MJ, McGuire P. How can neuroimaging facilitate the diagnosis and stratification of patients with psychosis? *Eur Neuropsychopharmacol* 2015;**25**:725–32.

118 Jackson R. TextHunter. Published Online First: 2 August 2014. doi:10.5281/zenodo.11122

119 Jackson R, Ball M, Patel R, *et al.* TextHunter - A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research. *Proc Am Med Informatics Assoc* 2014;:729–38. doi:10.13140/2.1.3722.9121

120 StataCorp. Stata Statistical Software: Release 12. *Coll Station TX StataCorp LP* 2011.

121 Gorrell G, Jackson R, Roberts A, *et al.* Finding Negative Symptoms of Schizophrenia in Patient Records. *Proc NLP Med Biol Work (NLPMedBio), Recent Adv Nat Lang Process (RANLP), Hissar, Bulg* 2013;:9–17.http://aclweb.org/anthology/W/W13/W13-5102.pdf

122 Nelson B, Yung AR. When things are not as they seem: detecting first-episode psychosis upon referral to ultra high risk (‘prodromal’) clinics. *Early Interv Psychiatry* 2007;**1**:208–11.

123 Blanchard JJ, Kring AM, Horan WP, *et al.* Toward the next generation of negative symptom assessments: the collaboration to advance negative symptom assessment in schizophrenia. *Schizophr Bull* 2011;**37**:291–9.

124 Neuhauser D, Diaz M. Quality improvement research: are randomised trials necessary? *Qual Saf Heal Care* 2007;**16**:77–80.

125 McCarney R, Warner J, Iliffe S, *et al.* The Hawthorne Effect: a randomised, controlled trial. *BMC Med Res Methodol* 2007;**7**:30.

126 South London and Maudsley NHS Foundation Trust. D-CRIS Product Description. 2015.http://www.slam.nhs.uk/media/239294/what_is_cris_-_product_description_v1_1.pdf

127 Howes OD, Weinstein S, Tabraham P, *et al.* Street slang and schizophrenia. *BMJ* 2007;**335**:1294.http://www.bmj.com/content/335/7633/1294.abstract

128 Kay SR, Fiszbein a, Opler L a. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr Bull* 1987;**13**:261–76.http://www.ncbi.nlm.nih.gov/pubmed/23434198

129 Patel R, Shetty H, Boydell J, *et al.* Factors affecting hospital admission following presentation to mental health services with psychosis. *Early Interv Psychiatry* 2014;**8**:159. doi:10.1111/eip.12199

130 EUropean network of national schizophrenia networks studying Gene-Environment Interactions (EU-GEI). Eur. Comm. 2015.http://www.eu-gei.eu/about-the-project/research-programme

131 Optimization of Treatment and Management of Schizophrenia in Europe (OPTiMiSE). Eur. Comm. 2015.http://www.optimisetrial.eu/

132 Translating neuroimaging findings from research into clinical practice (PSYSCAN). Eur. Comm. 2015.http://ec.europa.eu/research/health/medical-research/brain-research/projects/psyscan_en.html

133 Schizophrenia: Treatment Resistance and Therapeutic Advances (STRATA). Med. Res. Counc. 2015.https://www.kcl.ac.uk/ioppn/depts/ps/research/STRATA.aspx